

Novel Phylogenetic Approaches to Problems in Microbial Genomics

by

Lawrence A. David

Submitted to the Computational & Systems Biology Initiative
in partial fulfillment of the requirements for the degree of

PhD

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

© Massachusetts Institute of Technology 2010. All rights reserved.

Author
Computational & Systems Biology Initiative
August 26th, 2010

Certified by
Eric J. Alm
Assistant Professor
Thesis Supervisor

Accepted by
Chris Burge
Chair, Ph.D. Graduate Committee

| Report Documentation Page | | | | Form Approved OMB No. 0704-0188 | |
|--|------------------------------------|-------------------------------------|---|---|---------------------------------|
| Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. | | | | | |
| 1. REPORT DATE SEP 2010 | | 2. REPORT TYPE | | 3. DATES COVERED 00-00-2010 to 00-00-2010 | |
| 4. TITLE AND SUBTITLE Novel Phylogenetic Approaches to Problems in Microbial Genomics | | | | 5a. CONTRACT NUMBER | |
| | | | | 5b. GRANT NUMBER | |
| | | | | 5c. PROGRAM ELEMENT NUMBER | |
| 6. AUTHOR(S) | | | | 5d. PROJECT NUMBER | |
| | | | | 5e. TASK NUMBER | |
| | | | | 5f. WORK UNIT NUMBER | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139 | | | | 8. PERFORMING ORGANIZATION REPORT NUMBER | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | | | | 10. SPONSOR/MONITOR'S ACRONYM(S) | |
| | | | | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) | |
| 12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited | | | | | |
| 13. SUPPLEMENTARY NOTES | | | | | |
| 14. ABSTRACT Present day microbial genomes are the handiwork of over 3 billion years of evolution. Comparisons between these genomes enable stepping backwards through past evolutionary events, and can be formalized using binary tree models known as phylogenies. In this thesis, I present three new phylogenetic methods for gaining insight into how microbes evolve. In Chapter 1, I introduce the algorithm AdaptML, which uses strain ecology information to identify genetically- and ecologically-distinct bacterial populations. Analysis of 1000 marine Vibrionaceae strains by AdaptML finds evidence that niche adaptation may influence patterns of genetic differentiation in bacteria. In Chapter 2, I introduce the algorithm AnGST, which can infer the evolutionary history of a gene family in a chronological context. Analysis of 3968 gene families drawn from 100 modern day organisms with AnGST reveals genomic evidence for a massive expansion in microbial genetic diversity during the Archean eon and the gradual oxygenation of the biosphere over the past 3 billion years. Lastly, I introduce in Chapter 3 the algorithm GAnG, which can construct prokaryotic species trees from thousands of distinct gene trees. GAnG analysis of archaeal gene trees supports hypotheses that the Nanoarchaeota diverged from the last ancestor of the Archaea prior to the Crenarchaeota/Euryarchaeota split. | | | | | |
| 15. SUBJECT TERMS | | | | | |
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT Same as Report (SAR) | 18. NUMBER OF PAGES 126 | 19a. NAME OF RESPONSIBLE PERSON |
| a. REPORT unclassified | b. ABSTRACT unclassified | c. THIS PAGE unclassified | | | |

Novel Phylogenetic Approaches to Problems in Microbial Genomics

by

Lawrence A. David

Submitted to the Computational & Systems Biology Initiative
on August 26th, 2010, in partial fulfillment of the
requirements for the degree of
PhD

Abstract

Present day microbial genomes are the handiwork of over 3 billion years of evolution. Comparisons between these genomes enable stepping backwards through past evolutionary events, and can be formalized using binary tree models known as phylogenies. In this thesis, I present three new phylogenetic methods for gaining insight into how microbes evolve. In Chapter 1, I introduce the algorithm AdaptML, which uses strain ecology information to identify genetically- and ecologically-distinct bacterial populations. Analysis of 1000 marine *Vibrionaceae* strains by AdaptML finds evidence that niche adaptation may influence patterns of genetic differentiation in bacteria. In Chapter 2, I introduce the algorithm AnGST, which can infer the evolutionary history of a gene family in a chronological context. Analysis of 3968 gene families drawn from 100 modern day organisms with AnGST reveals genomic evidence for a massive expansion in microbial genetic diversity during the Archean eon and the gradual oxygenation of the biosphere over the past 3 billion years. Lastly, I introduce in Chapter 3 the algorithm GAnG, which can construct prokaryotic species trees from thousands of distinct gene trees. GAnG analysis of archaeal gene trees supports hypotheses that the Nanoarchaeota diverged from the last ancestor of the Archaea prior to the Crenarchaeota/Euryarchaeota split.

Thesis Supervisor: Eric J. Alm
Title: Assistant Professor

Acknowledgments

Funding

I am deeply grateful for the research support provided by:

- National Defense Science & Engineering Graduate Fellowship (DoD)
- Whitaker Health Sciences Fund Fellowship (MIT)

Personal

Many people have deliberately or unwittingly helped me complete my doctoral work. My words of gratitude below are only a small installment against a lifetime of welcome debt.

Foremost, I would like to thank my advisor, Eric Alm, who gave me the freedom to work on the problems I found interesting and the scientific training to solve them. His unique enthusiasm over these past years has helped fuel my research and provided me with personal joys that I will always cherish.

I have been also remarkably fortunate to have joined a group of wonderful colleagues here at MIT. Mark Smith, Chris Smillie, Greg Fournier, Sarah Preheim, Ines Baptista, Matt Blackburn, Yonatan Friedman, Alexandra Konings, Claudia Bauer, and Albert Wang have been a source of thoughtful discussion and much needed levity. In particular, the old guard of Arne Materna, Sean Clarke, Sonia Timberlake, and Jesse Shapiro is enshrined in some of my fondest memories of graduate school. Classmates in my department, especially Robin Friedman, Charles Lin, and Michelle Chan, have been inspirational young-scientists who I have looked to for advice and motivation.

A host of caring individuals enabled me to reach the closing stages of my graduate career. Sheila Frankel, Darlene Strother, Darlene Ray, James Long, and Bonnie Lee Whang supplied administrative and moral support. My graduate committee, whose membership has included Ed DeLong, Drew Endy, Manolis Kellis, Dianne Newman, Howard Ochman, (and unofficially Martin Polz), generously set aside time to provide me with professional advice and letters of recommendation. [greg] ensured that my cluster jobs ran smoothly and provided crucial UNIX humor.

I also acknowledge my friends and family, who conspired to make graduate school the shortest five years of my life. Within the Parsons laboratory, David Gonzalez-Rodriguez, Piyatida Hoisungwan, Marcos, and Crystal Ng, unfailingly made me excited to come to lab each morning. MacGregor's F-entry has been my adopted undergraduate family, and I hold dear my memories of living among them. My sister has continuously given me her love, and my father and mother have selflessly worked to give me the opportunities I presently enjoy. Finally, my wife Christina has simply been the nicest person I've ever met.

Contents

| | | |
|-----|--|-----|
| I | Introduction | 5 |
| II | Contributions | 13 |
| 1 | Resource partitioning and sympatric differentiation among closely related bacterioplankton | 15 |
| 2 | Rapid evolutionary innovation during an Archean Genetic Expansion | 51 |
| 3 | Building prokaryotic species trees from thousands of gene trees | 92 |
| III | Conclusion | 113 |
| IV | Bibliography | 116 |

Part I

Introduction

Overview

Present day microbial genomes are the handiwork of over 3 billion years of evolution. Comparisons between these genomes enable stepping backwards through evolutionary history – genetic features present across a wide diversity of genomes likely arose more anciently than features found in subsets of related genomes. This intuition is formalized using binary tree models of sequence evolution known as phylogenetic trees. Phylogenies propose a series of ancestral sequence divergence events that explain the similarity of extant sequences. These trees can in turn be used to build models for the evolution of organismal phenotypes, such as preferred environment or lifestyle. However, prokaryotes’ capacity for horizontal gene transfer (HGT) can require regions of the same genome to be associated with different phylogenetic trees, and ultimately obscure which phylogenetic tree best represents overall genome evolution.

In this thesis, I present three novel phylogenetic approaches for inferring microbial evolutionary history through the comparison of gene sequences. The remainder of Part I briefly describes the research context in which I developed: AdaptML, an algorithm for detecting signatures of ecological adaptation influencing bacterial genetic differentiation; and AnGST, an algorithm for inferring the series of HGT, gene duplication, and gene loss events that gave rise to a gene family. I go on in Chapter 1 of Part II to use AdaptML to identify genetically- and ecologically-distinct clusters of *Vibrionaceae* coexisting in a marine environment. In Chapter 2, I use AnGST to infer patterns in microbial genome evolution over the past 3.8 billion years. I use AnGST again in Chapter 3 in the development of GAnG, a new method for constructing prokaryotic species trees from thousands of gene trees. I conclude this thesis in Part III with a summary of the chapters and a brief discussion of ongoing and future work with AdaptML, AnGST, and GAnG.

Detecting relationships between genetic and ecological differentiation in bacteria

Distinct groups of closely-related bacteria, or phylogenetic clusters, are a recurring pattern of genetic differentiation among bacterial isolate housekeeping genes [1–3]. Ecological adaptation is suspected of playing a role in cluster formation [4]. According to the ecotype model, genetically-distinct bacterial populations form when bacterial populations adapt to an ecological niche and are repeatedly purged of genetic variation through periodic selection events [5]. However, a theoretical study has shown that genetically distinct sub-populations can form under a neutral model that either prohibits recombination, or simulates high within-cluster recombination [6]. Alternatively, a recent phylogenetic analysis of eight sequenced *Vibrio* isolates has found evidence for a combined model featuring both ecological adaptation and neutral processes contributing to genetic differentiation. Under this model, the introduction of niche-adaptive alleles initially erodes sympatry in a bacterial population. Reduced gene flow between niche-adapted bacteria and the remaining population subsequently yields genetically-distinct subgroups [7]. Ultimately, if niche adaptation drives the formation of genetically-distinct bacterial groups, members of each group should inhabit a common niche. Mathematical models capable of identifying both genetically- and ecologically-cohesive bacterial groups can thus be used to help resolve the role of ecological adaptation in the genetic differentiation of bacteria.

Several existing statistical methods, such as the Fst test, the P test, and Unifrac, can evaluate the null hypothesis that phylogenetic clusters do not exhibit distinct ecological associations. These tests assume that bacterial sequences are annotated with ecological metadata describing the environment each sequence was harvested from. The Fst test compares the genetic diversity among bacteria annotated as sharing the same environment to the genetic diversity measured across all sampled sequences. Low genetic diversity within a particular environment, coupled with high genetic diversity between environments, is evidence for rejection of the null hypothesis of no association between genetic clustering and bacterial ecology [8]. Alternatively,

the P test builds a phylogeny of strain sequences, labels leaves by their environmental association, and uses a parsimony model to infer the number of times ancestral strains on the tree changed environmental associations. Low parsimony scores are evidence for rejecting the null hypothesis [8]. Lastly, the Unifrac model combines elements of both the Fst and P test, utilizing genetic distances and strain tree topology to test the relationship between strain genetic clustering and associated environment [9].

One weakness, however, of the Fst, P, and Unifrac statistics is their potential for erroneously reporting no association between genetic clustering and ecology when the ecological forces driving cluster formation are unmeasured or improperly annotated. For example, consider a bacterial sequence cluster caused by adaptation to conditions between 20°-30°C. An association between this cluster and ecology would go unrecognized by Fst, P, or Unifrac analyses if temperature data was not collected, or if temperature data were discretized into only two ranges: $< 25^{\circ}\text{C}$ and $\geq 25^{\circ}\text{C}$. Thus, these statistics may not be appropriate for analyzing the evolution of bacteria whose niche composition is unknown or highly uncertain, as environmental parameters describing these bacteria’s niche may not have been measured.

Another inference algorithm, Ecotype Simulation (ES), can identify genetically- and ecologically-distinct clusters in a manner insensitive to how ecological parameters are measured [10]. ES finds ecotypes by fitting a maximum likelihood model onto a gene phylogeny. This model estimates the rates of ecotype formation, periodic selection, and genetic drift, as well as the total number of ecotypes present. Identified ecotypes can subsequently be analyzed using ecological measurements and multivariate statistics in order to confirm that niche-adaptation has taken place and identify environmental parameters that define the niche. Recent application of this approach discovered ecotypes among *Bacillus* strains sampled from Death Valley, CA, which could be distinguished by adaptation to solar exposure and soil texture [11]. One drawback to the ES algorithm, however, is that it cannot detect nascent ecotype formation events that have not yet undergone multiple series of periodic selections.

In Chapter 1, I present a new method named AdaptML, which uses a maximum likelihood model to identify genetically- and ecologically-coherent clusters of

bacterial strains. This model explicitly combines genetic information embedded in sequence-based phylogenies with environmental sampling data. Recent niche adaptation events, characterized by ecologically coherent clusters with minimal genetic distinction from a parent clade, can be captured by the model. Although AdaptML cannot detect ecological associations with unmeasured environmental parameters, the algorithm can account for environmental parameter discretization schemes that would generally confound previous methods for detecting ecological associations. To do this, I introduce the model concept of a “habitat.” Habitats are characterized by discrete probability distributions describing the likelihood that a strain adapted to a habitat will be sampled from a given ecological state (e.g. at a particular location in an estuary). Habitats are not defined *a priori* but rather learned directly from the sequence and ecological data using an Expectation Maximization routine. Once habitats are defined, I learn a maximum likelihood model for the evolution of habitat association on the tree. Randomization experiments can be used to determine which sequence clusters show a statistically-significant association with a given habitat.

Inferring the evolutionary history of microbial gene families

Microbial genomes do not evolve solely by point mutation [12, 13]. Comparison of gamma-proteobacterial genomes suggests gene loss events eliminated thousands of genes from the ancestor of the *Buchnera* following its adoption of an endosymbiotic lifestyle [14]. Genes can be gained, via either the duplication of small regions of the genome [15], or via the duplication of the entire genome itself, as has been shown for yeast [16]. Gene gain is also possible via HGT and is a well-known source of genomic diversity among the prokaryotes [12]. Cases of HGT have also been identified between eukaryotes [17, 18] and even from bacteria to animals [19, 20]. Models that can infer when genes have undergone loss, duplication, or HGT, and when genes have been vertically inherited, are necessary for understanding the relative contribution of these four mechanisms to genome evolution.

Algorithms for inferring the evolutionary history of gene families vary according to their reliance on phylogenetic models and how they account for gene gain events (Table 1). Phylogeny-free methods utilize features such as GC-bias to detect xenologous genes [21, 22], or within-genome BLAST searches to find evidence for past duplication events [23]. More complex approaches, known as presence-absence models, construct a phylogeny of sampled species and identify which leaves on the tree are represented in a gene family of interest. Parsimony algorithms can then be used to identify a set of ancestral gene duplication, gene loss, or HGT events to explain the observed pattern of gene presence and absence on the species tree [24–27]. However, presence/absence algorithms may underestimate the amount of HGT in a gene family history, since frequent HGT events can produce presence/absence patterns similar to those caused by gene birth at a deep node, followed by vertical descent. More sensitive models capable of differentiating between these scenarios utilize gene sequence information, in addition to a species tree. Quartet methods quantify how strongly quartets of orthologous genes support each of the three possible 4-taxon trees representing their evolutionary history [28, 29]. Quartets that strongly support topologies discordant

| Model | Species tree | Gene tree | Unc. gene trees | Finds HGT | Finds dup. | Refs. |
|------------------------------|--------------|-----------|-----------------|-----------|------------|----------|
| GC-bias | No | No | - | Yes | No | [21, 22] |
| BLAST-hits | No | No | - | Yes | Yes | [23] |
| Presence/absence | Yes | No | - | Yes | Yes | [24–27] |
| Quartet mapping | Partial | Partial | Yes | Yes | No | [28, 29] |
| Parsimony reconciliation | Yes | Yes | No | Yes | Yes | [30, 31] |
| Probabilistic reconciliation | Yes | Yes | Yes | Yes | No | [33, 34] |

Table 1: Selection of existing models used to infer gene family evolutionary histories: Models are characterized by their explicit usage of species trees and gene trees, their consideration of gene tree uncertainty, and their ability to detect HGT and duplication events. Note that only References [27, 31] can find HGT and duplication events simultaneously.

with the expected species tree are evidence for HGT within the gene family. More elaborate “reconciliation” models compare full gene and species trees in order to infer a precise phylogenetic location for each inferred evolutionary event. Parsimony reconciliation models [30, 31], however, will infer spurious events if phylogenetic construction errors are present in the gene tree [32]. Newer probabilistic reconciliation algorithms have been developed to deal with these potential inaccuracies [33, 34].

Gene family evolutionary history models can also be partitioned according to whether they account for gene gain using duplication or HGT events. With the exception of Snel and Charleston’s algorithms [27, 31], evolutionary history models usually account for only one of these two events. The specificity of these models may be caused by self-reinforcing biases associated with the expected modes of eukaryotic and prokaryotic genome evolution. The relative rarity of reported HGT events among eukaryotes, compared to duplication events, likely encourages analyses of eukaryotic genome evolution using tools specialized only to detect gene duplications. By contrast, recognition of how HGT can accelerate prokaryotic adaptation and blur species lines has probably reduced interest in broad surveys of potential prokaryotic duplication. Exceptions to this proposed bias among prokaryotic studies do exist, however, as

Gevers et al. cataloged gene duplications in 106 bacterial genomes [23] and Snel searched for both HGT and duplication among 17 archaeal and bacterial genomes [27]. Increasing examples of HGT among eukaryotes [17, 18] are also fueling new interest in systematically searching for HGT across the eukaryotes [35]. Bias against the creation of models that account for both HGT and duplication is also likely due to issues of model complexity. In certain scenarios, gene duplication and gene loss can produce gene tree topologies similar to those yielded by HGT [36]. A combined HGT/duplication inference model must be capable of recognizing this scenario and proposing plausible HGT and duplication scenarios. Moreover, a combined model requires defining a metric to choose which of these scenarios is preferable.

In Chapter 2, I present a new reconciliation method for inferring a set of gene loss, gene duplication, and HGT events that explain topological incongruities between a species tree and a gene tree. I named this algorithm the Analyzer of Gene & Species Trees, or AnGST. AnGST was inspired by a gene family evolution model originally designed for problems in biogeography and the inference of gene duplication and gene loss events [30]. Also referred to as a host-parasite model, this approach seeks to infer which ancestral genome (the host) on the reference tree possessed each ancestral gene copy (the parasite). AnGST employs a generalized parsimony framework in order to choose when duplication events should be inferred instead of HGT scenarios. This framework assigns scores to each type of evolution event and returns the evolutionary history with the lowest overall score. AnGST can further minimize reconciliation scores by reconciling multiple gene tree bootstraps simultaneously and combining their lowest scoring subtrees into a single chimeric gene tree. This bootstrap amalgamation step reduces the opportunity for poorly resolved gene tree subtrees to cause the spurious inference of evolutionary events.

Part II

Contributions

Chapter 1

Resource partitioning and sympatric differentiation among closely related bacterioplankton

Dana E. Hunt*, Lawrence A. David*, Dirk Gevers, Sarah P. Preheim,
Eric J. Alm, Martin F. Polz

*These authors contributed equally to this work.

This chapter is presented as it originally appeared in *Science* **320**, 1081 (2008).
Corresponding Supplementary Material is appended.

Chapter 1

Resource partitioning and sympatric differentiation among closely related bacterioplankton

Identifying ecologically differentiated populations within complex microbial communities remains challenging, yet is critical for interpreting the evolution and ecology of microbes in the wild. Here we describe spatial and temporal resource partitioning among *Vibrionaceae* strains coexisting in coastal bacterioplankton. A quantitative model (AdaptML) establishes the evolutionary history of ecological differentiation, thus revealing populations specific for seasons and life-styles (combinations of free-living, particle, or zooplankton associations). These ecological population boundaries frequently occur at deep phylogenetic levels (consistent with named species); however, recent and perhaps ongoing adaptive radiation is evident in *Vibrio splendidus*, which comprises numerous ecologically distinct populations at different levels of phylogenetic differentiation. Thus, environmental specialization may be an important correlate or even trigger of speciation among sympatric microbes.

Microbes dominate biomass and control biogeochemical cycling in the ocean, but we know little about the mechanisms and dynamics of their functional differentiation in the environment. Culture-independent analysis typically reveals vast microbial diversity, and although some taxa and gene families are differentially distributed among environments [37, 38], it is not clear to what extent coexisting genotypic diversity can be divided into functionally cohesive populations [37, 39]. First, we lack broad surveys of nonpathogenic free-living bacteria that establish robust associations of individual strains with spatiotemporal conditions [40, 41]; second, it remains controversial what level of genetic diversification reflects ecological differentiation. Phylogenetic clusters have been proposed to correspond to ecological populations that arise by neutral diversification after niche-specific selective sweeps [5]. Clusters are indeed observed among closely related isolates (e.g., when examined by multilocus sequence analysis) [4] and in culture-independent analyses of coastal bacterioplankton [42]. Yet recent theoretical studies suggest that clusters can result from neutral evolution alone [6], and evidence for clusters as ecologically distinct populations remains sparse, having been most conclusively demonstrated for cyanobacteria along ocean-scale gradients [43] and in a depth profile of a microbial mat [44]). Further, horizontal gene transfer (HGT) may erode the ecological cohesion of clusters if adaptive genes are transferred [45], and recombination can homogenize genes between ecologically distinct populations [46]. Thus, exploring the relationship between phylogenetic and ecological differentiation is a critical step toward understanding the evolutionary mechanisms of bacterial speciation [6].

In this study, we investigated ecological differentiation by spatial and temporal resource partitioning in coastal waters among coexisting bacteria of the family *Vibrionaceae*, which are ubiquitous, metabolically versatile heterotrophs [47]. The coastal ocean is well suited to test population-level effects of microhabitat preferences, because tidal mixing and oceanic circulation ensure a high probability of migration, reducing biogeographic effects on population structure. In the plankton, heterotrophs may adopt alternate ecological strategies: exploiting either the generally lower concentration but more evenly distributed dissolved nutrients or attaching to and degrading

small suspended organic particles, originating from algal exopolysaccharides and detritus [39]. Bacterial microhabitat preferences may develop because resources are distributed on the same scale as the dispersal range of individuals, due to turbulent mixing and active motility [48]. Of potential microhabitats, particles represent abundant but relatively short-lived resources, as labile components are rapidly utilized (on time scales of hours to days) [49, 50], implying that particle colonization is a dynamic process. Moreover, particulate matter may change composition with macroecological conditions (such as seasonal algal blooms). Zooplankton provide additional, more stable microhabitats; vibrios attach to and metabolize chitinous zooplankton exoskeletons [51, 52] but may also live in the gut or occupy niches specific to pathogens. The extent to which microenvironmental preferences contribute to resource partitioning in this complex ecological landscape remains an important question in microbial ecology [53].

We aimed to conservatively identify ecologically coherent groups by examining distribution patterns of *Vibrionaceae* genotypes among free-living and associated (with suspended particles and zooplankton) compartments of the planktonic environment under different macroecological conditions (spring and fall) (Figs. 1.3 & 1.5). Because the level of genetic differentiation at which ecological preferences develop is not known, we focused on a range of relationships (0 to 10% small subunit ribosomal RNA (rRNA) divergence) among co-occurring vibrios [54]. Particle-associated and free-living cells were separated into four consecutive size fractions by sequential filtration (four replicate water samples, each subsampled with at least four replicate filters per size fraction); each fraction contained organisms and dead organic material of different origins (detailed in the supporting online material [SOM]; Section 1.2). For simplicity, we refer to these fractions as enriched in zooplankton (≥ 63 μm), in large (5 to 63 μm) and small (1 to 5 μm) particles, and in free-living cells (0.22 to 1 μm) (Fig. 1.5B). The 1- to 5- μm size fraction was somewhat ambiguous, probably containing small particles as well as large or dividing cells; however, it provided a firm buffer between obviously particle-associated (>5 μm) and free-living (<1 μm) cells. *Vibrionaceae* strains were isolated by plating filters on selective media, previously

shown by quantitative polymerase chain reaction to yield good correspondence between genotypes recovered in culture and those present in environmental samples [54]. Roughly 1000 isolates were characterized by partial sequencing of a protein-coding gene (*hsp60*). To obtain added resolution, between one and three additional gene fragments (*mdh*, *adk*, and *pgi*) were sequenced for over half of the isolates (SOM), including *V. splendidus* strains, the most abundant group [54].

Our rationale for testing environmental associations grows out of the following considerations. First, as in most ecological sampling, the true habitats or niches are unknown and can only be observed as projections onto the sampling dimensions (“projected habitats”). Thus, associations can be detected as distinct distributions of groups of strains if habitats/niches are differentially apportioned among samples. Second, the lack of an accepted microbial species concept implies that it is imprudent to use any measure of genetic relationships to define a priori the populations whose environmental association should be assessed. Therefore, we first tested the null hypothesis that there is no environmental association across the phylogeny of the strains. We then refined such estimates by developing a new model to simultaneously identify populations and their projected habitats. Finally, these model-based results were tested with nonparametric empirical statistics.

The initial null hypothesis of no association between phylogeny and ecology is strongly rejected (seasons: $p < 10^{-79}$; size fractions: $p < 10^{-49}$) by comparing the parsimony score of observed environments on the tree to that expected by chance [55] (SOM), confirming the visual impression of differential patterns of clustering among seasons and size fractions (Fig. 1.1A). This result is robust toward uncertainty in the phylogeny, which should diminish but not strengthen associations, and is confirmed by introducing additional uncertainty in the phylogeny (Fig. 1.6). The observed overall association with season and size fraction therefore suggests that water-column vibrios partition resources, but neither provides insights into the phylogenetic bounds of populations or the composition of their habitats.

We therefore developed an evolutionary model (AdaptML) to identify populations as groups of related strains sharing a common projected habitat, which reflects

their relative abundance in the measured environmental categories (size fractions and seasons) (SOM). In practice, the model inputs are the phylogeny, season, and size fraction of the strains. It then maps changes in environmental preference onto the tree by predicting projected habitats for each extant and ancestral strain in the phylogeny. Although similar in spirit to existing parsimony, likelihood, and Bayesian methods, which map ancestral states onto trees [56], the model accounts for the complexities and uncertainties of environmental sampling. First, projected habitats can span multiple sampling dimensions to account for complex life cycles (such as time spent in multiple true habitats) and problems inherent in environmental sampling: Discrete samples rarely equate to true habitats, and true habitats are frequently misplaced among their typical sample categories (for example, zooplankton fragments may also be found in smaller size fractions). Second, projected habitats can span multiple phylogenetic clusters to allow for the possibility that clusters may arise neutrally or that the relevant parameters differentiating them ecologically have not been measured.

Briefly, AdaptML builds a hidden Markov model for the evolution of habitat associations: Adjacent nodes on the phylogeny transition between habitats according to a probability function that is dependent on branch length and a transition rate, which is learned from the data (SOM) (Fig. 1.7). Subsequently, we optimize the model parameters (the transition rate and the composition of each projected habitat) to maximize the likelihood of the observed data. Finally, we use a simple ad hoc rule for reducing noninformative parameters: We merge habitats that converge to similar distributions (simple correlation of distribution vectors $>90\%$) during the model-fitting procedure (SOM). This reproducibly identified six nonredundant habitats for the observed data set (H_A to H_F in Figs. 1.1B and 1.9). Moreover, the algorithm acts conservatively, as suggested by two tests. First, the model did not overfit the data when there was no ecological signal present: When the environments were shuffled, only a single generalist habitat (evenly distributed over all size fractions and seasons) was recovered. Second, when simulated habitats were used to generate environmental assignments, the model usually identified a number of habitats equal to or less than

the true number present (Fig. 1.10).

The analysis suggests that a single bacterial family coexisting in the water column resolves into a striking number of ecologically distinct populations with clearly identifiable preferences (habitats). The algorithm identified 25 populations, associated with one of the six habitats defined by distinct distributions of isolates over seasons and size fractions (Fig. 1.1 and Fig. 1.11). Most clusters have a strong seasonal signal; interestingly, two pairs of highly similar habitats are observed in both seasons (Fig. 1.1B). The first of the habitat pairs corresponds to populations occurring both free-living and on particles but lacking zooplankton-associated isolates (H_B and H_C); the second indicates a preference for zooplankton and large particles (H_E and H_F) (Fig. 1.1B). The remaining two habitats were season-specific. Habitat H_A combines all primarily free-living populations in the fall, whereas habitat H_D identifies a second particle- and zooplankton-associated group in spring, but unlike H_E and H_F it has a higher proportion of large particles and maps onto a single small group (G25) (Fig. 1.1). However, we cannot place high confidence in the absence of the free-living habitat in the spring, because relatively few strains were recovered from that fraction. Moreover, the distribution of individual populations among seasons and size fractions varies considerably, with remarkably narrow preferences for some populations whereas others are more broadly distributed. For example, *V. ordalii* (G3) is almost exclusively free-living in both seasons, whereas *V. alginolyticus* (G5) has a significant representation in both zooplankton and free-living size fractions but occurs exclusively in the fall (Fig. 1.1, A and B). The sequences of three additional genes for *V. alginolyticus* isolates were identical, arguing against misidentification due to recombination or additional population substructuring. Similarly, there was good agreement when two different gene phylogenies (*hsp60* and *mdh*) were used to identify habitats for *V. splendidus* (Fig. 1.12), although fewer habitats were identified using the *mdh* tree, most likely because it is less well-resolved. Overall, across all vibrios sampled, association with the zooplankton-enriched and free-living fractions dominated, and although several populations contain particle-associated isolates, only a few appear to be specifically particle-adapted. Because vibrios are generally regarded as particle

and zooplankton specialists [47], this observed partitioning offers new insight into their ecology.

Thus, in spite of the highly variable conditions of the water column, populations appear to finely partition resources, especially because our habitat estimates are conservative, as clusters occupying the same habitat may be differentiated along additional (unobserved) resource axes. For example, different zooplankton-associated groups may be host- or body region-specific, and the strong seasonal signal of most clusters may be due to a variety of factors; however, temperature is a likely candidate because it has so far arisen as the strongest correlate of microbial population changes both over a seasonal cycle [57] and along ocean-scale gradients [43]. Finally, populations, which appear unassociated in our study, may be true generalists with respect to the resource space sampled or may be adapted to environments not sampled in this study, such as animal intestines or sediments [47]. Despite these uncertainties, the observed strong partitioning among associated and free-living clusters may have important implications for population biology in the bacterioplankton. As recently suggested [6], for attached bacteria, the effective population size (N_e) may be considerably smaller than the census size because colonization serves as a population bottleneck, whereas in free-living clusters, N_e may be closer to the census size. Although computing the true magnitude of N_e in microbial populations remains controversial [58], it is an important parameter that determines the relative strength of selection and drift. Thus, attached and free-living populations may evolve under different constraints [6].

The phylogenetic structure of populations also provides insights into the history of habitat switches. Deeply branching populations may have remained associated with habitats over long evolutionary time, and shallow branches may have diversified more recently (Fig. 1.1, A and B). These stable habitat-associated clusters roughly correlate to named species within the *Vibrionaceae*. For example, *V. ordalii* (G3) and *Enterovibrio norvegicus* (G2) both represent clusters without close relatives containing > 50 isolates, which are overwhelmingly predicted to follow primarily free-living (H_A) and free-living/particle-associated lifestyles (H_C), respectively (Fig. 1.1A). On

the other hand, some very closely related clusters are associated with different habitats; *V. splendidus*, which is composed of strains that are $\sim 99\%$ identical in rDNA gene sequence [54], differentiates into 15 microdiverse habitat-associated clusters, of which one is distributed roughly evenly among both seasons, and 9 and 5 predominantly occur in spring and fall, respectively. Thus, *V. splendidus* appears to have ecologically diversified, possibly by invading new niches or partitioning resources at increasingly fine scales.

Recent or perhaps ongoing radiation by sympatric resource partitioning is most strongly suggested for two nested clusters within *V. splendidus*, where groups of strains differing by as little as a single nucleotide in *hsp60* display distinct ecological preferences (Fig. 1.1A, insets, and Fig. 1.3). These strains were isolated from multiple independent samples and thus do not represent clonal expansion, suggesting that this may reflect a true habitat switch; nonetheless, homologous recombination could also move alleles between distantly related, ecologically distinct clusters, creating spurious phylogenetic relationships, which can be detected by comparison with other genes. Multilocus sequence analysis shows that for nested cluster I, a close relationship was artificially created because *hsp60* gene phylogeny is discordant with three other genes (Fig. 1.2). However, this still represents a habitat switch, just at a slightly larger sequence distance, as I.A is nested within the much larger G16 cluster in both the *hsp60* and the *mdh-pgi-adk* phylogenies. For the second nested cluster, the three additional genes confirm partial separation of the subclusters II.A and II.B by a single base pair difference in one of the genes, whereas the other genes consist of identical alleles. This reinforces the idea that subcluster II.A is not incorrectly grouped because of recombination, despite its distinct ecological affiliation (Fig. 1.2). In combination, these data support the idea that there is ecological differentiation among recently diverged genotypes and show that such changes might be recognized in protein-coding genes as soon as they accumulate (neutral) sequence changes.

How might adaptation to a new habitat relate to speciation, the generation of distinct clusters of closely related bacteria? Mathematical modeling has recently shown that the dynamics of speciation depend on the ratio of homologous recombination to

mutation rates (r/m) [6]. When this ratio per allele exceeds ~ 1 , populations transition from essentially clonal to sexual, with the major consequence that selection is probably required for the formation of clusters [6]. Our preliminary multilocus sequence analysis on a set of strains with similar taxonomic composition suggests that their r/m is well above that threshold. Thus, our observations of habitat separation for highly similar but clearly distinct genotypes suggest that ecological selection may have triggered phylogenetic differentiation. A plausible mechanism is that differential distribution among habitats (possibly caused by few adaptive loci) is sufficient to depress gene flow between associated genotypes [6, 59]. Consequently, mutations will no longer be homogenized but instead accumulate within specialized populations, even for ecologically neutral genes. Over time, genetic isolation may increase because homologous recombination rates decrease log-linearly with sequence distance [60]. We detected associations with different habitats among sister clades over a wide range of phylogenetic distances, possibly representing populations at various stages of differentiation (Fig. 1.1A). Although we cannot determine whether clusters represent transiently adapted populations or nascent species, our observations of differential distributions of genotypes suggest that there exists a small-scale adaptive landscape in the water column allowing the initiation of (sympatric) speciation within this community.

Although it has recently been suggested that microbial lineages remain specific to macroenvironments over long evolutionary times [61], this study demonstrates switches in ecological associations within a bacterial family coexisting in the coastal ocean. In the *V. splendidus* clade, speciation could be ongoing, but the divergence between most other ecologically defined groups appears large. This is consistent with our previous suggestion that rRNA gene clusters, which are roughly congruent with the deeply divergent protein-coding gene clusters detected here, represent ecological populations [42]. However, the example of *V. splendidus* highlights the fact that using marker genes to assess community-wide diversity may not capture some ecological specialization. Moreover, different groups of organisms could evolve under different constraints, and the mechanisms suggested here apply to the invasion of new habitats

and are thus different from (but compatible with) the widely discussed niche-specific selective sweeps [10]. Why *V. splendidus* appears to have radiated recently into new habitats whereas other groups appear to be more constant is not known but may be related to its high heterogeneity in genome architecture [54]. This could indicate a large (flexible) gene pool that, if shared by horizontal gene transfer, gives rise to large numbers of ecologically adaptive phenotypes. It will therefore be important to compare whole genomes within recently ecologically diverged clusters to identify specific changes leading to adaptive evolution.

1.1 Figures

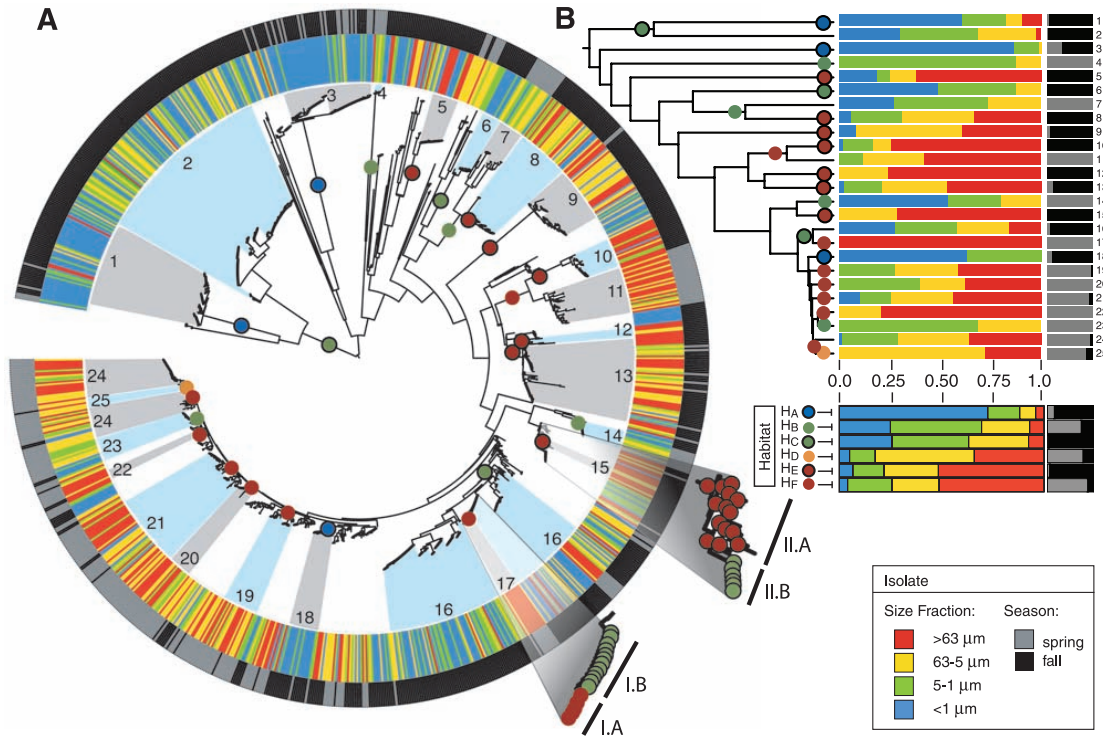


Figure 1.1: Season and size fraction distributions and habitat predictions mapped onto *Vibronaceae* isolate phylogeny inferred by maximum likelihood analysis of partial *hsp60* gene sequences. Projected habitats are identified by colored circles at the parent nodes. (A) Phylogenetic tree of all strains, with outer and inner rings indicating seasons and size fractions of strain origin, respectively. Ecological populations predicted by the model are indicated by alternating blue and gray shading of clusters if they pass an empirical confidence threshold of 99.99% (see SOM for details). Bootstrap confidence levels are shown in Fig. 1.14. (B) Ultrametric tree summarizing habitat-associated populations identified by the model and the distribution of each population among seasons and size fractions. The habitat legend matches the colored circles in (A) and (B) with the habitat distribution over seasons and size fractions inferred by the model. Distributions are normalized by the total number of counts in each environmental category to reduce the effects of uneven sampling. The insets at the lower right of (A) show two nested clusters (I.A and I.B and II.A and II.B) for which recent ecological differentiation is inferred, including habitat predictions at each node. The closest named species to numbered groups are as follows: G1, *V. calviensis*; G2, *Enterovibrio norvegicus*; G3, *V. ordalii*; G4, *V. rumoiensis*; G5, *V. alginolyticus*; G6, *V. aestuarianus*; G7, *V. fischeri/logei*; G8, *V. fischeri*; G9, *V. superstes*; G10, *V. penaeicida*; G11 to G25, *V. splendidus*.

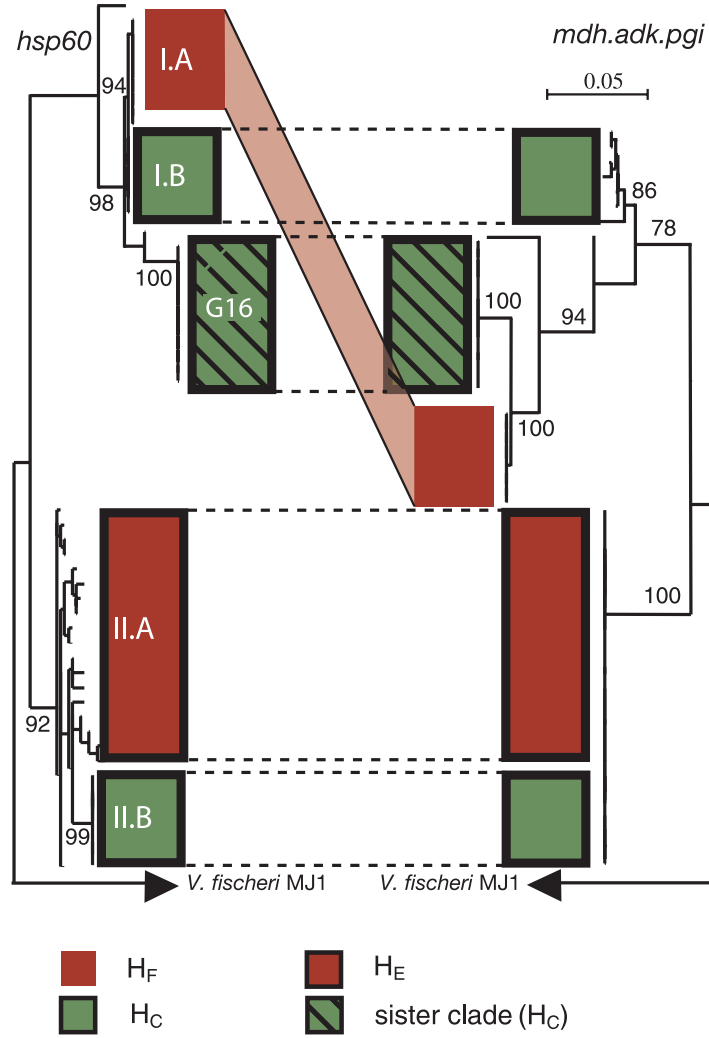


Figure 1.2: Multilocus sequence analysis of nested clusters (IA and IB and IIA and IIB) with differential habitat association by comparison of partial *hsp60* (left) and concatenated partial *mdh*, *adk*, and *pgi* (right) gene phylogenies. Habitat predictions (indicated by colored boxes) and the numbering of clusters correspond to Fig. 1.1. Scale bar is in units of nucleotide substitutions per site.

1.2 Supplementary Material

1.2.1 Sampling rationale

To investigate partitioning of *Vibrionaceae* strains in the water column, we examined their distribution among the free-living and associated (with particles and zooplankton) fractions of the bacterioplankton community at two time points. This was achieved by sequential filtration with decreasing pore size cutoffs and subsequent cultivation on *Vibrio* selective media (Fig. 1.5B). Here, we give additional details on sampling protocols and rationale supplementing the overview given in the main text.

Filtration is commonly used in oceanography to separate particle-associated and free-living populations by retention of particles on filters, although the filter size cut off for collecting particle-attached bacteria has varied in past studies between 0.8 and 10 μm [62–65]. To obtain higher ecological resolution, we used sequential filtration since alternate types of particulate organic matter and organisms (e.g., phytoplankton, zooplankton) will have distribution maxima in different size fractions thus enabling differentiation of associated bacterial genotypes.

We collected a total of four size fractions with different expected composition of particles and organisms (Fig. 1.5B). The largest fraction ($\geq 63 \mu\text{m}$) was visually enriched in zooplankton and detrital material (e.g., pieces of macroalgae, terrestrial plant material); however, large gelatinous material [frequently part of marine snow, which represents particles $>0.5 \text{ mm}$ [66]] was likely not collected since it is disrupted by the pressure on the plankton nets used for collection. All other fractions were collected by gravity rather than vacuum filtration to minimize disruption of fragile particles. The large particle fraction ($63\text{--}5 \mu\text{m}$) likely contains zooplankton fecal pellets, dead and living algae, and other detritus. The composition of the $5\text{--}1 \mu\text{m}$ size fraction is somewhat ambiguous since it may contain both cells attached to very small particles as well as large or dividing cells; however, it provides a firm buffer between obviously particle-attached ($>5 \mu\text{m}$) and free-living ($<1 \mu\text{m}$) cells. Particulate material in this size range may include small algae, bacterial cell walls, as well as fragments of larger particles, which have broken apart; nonetheless, the small size of

such particles are unlikely to sustain a resident bacterial population. Free-living bacteria, observed in the 1-0.22 μm size fraction, likely live on dissolved organic matter produced by living algae, cell lysis and the dissolution of particles.

1.2.2 Sample collection

Coastal ocean water samples were collected at high tide on the marine end of the Plum Island Estuary (NE Massachusetts) (Fig. 1.5A) on two days representing spring (4/28/06) and fall (9/6/06) conditions in the coastal ocean. Nutrient concentrations, water temperature and chlorophyll levels were measured on both sampling dates (Fig. 1.3).

Two replicate samples of the largest size fraction (enriched in zooplankton) were collected by filtering ~ 100 L each through a 63 μm plankton net, which was subsequently washed with sterile seawater (Fig. 1.5B). Particle-associated and free-living bacterial populations were collected from quadruplicate water samples, which were independently 2 pre-filtered through the 63 μm plankton net (to remove the zooplankton-enriched fraction) into 4 L nalgene bottles (Fig. 1.5B). For each bottle, water was sequentially filtered through 5, 1 and 0.22 μm pore size filters, collecting at least four replicate filters per size fraction. To avoid disruption of fragile particles, the 63-5 and 5-1 μm fractions were collected on polycarbonate membrane filters (Sterlitech) using gravity filtration followed by washing with 5 ml of sterile (0.22 μm -filtered and Tindalized) seawater to remove free-living bacteria that might have been retained on the filter. The <1 μm fraction containing free-living bacteria was collected on 0.22 μm Supor-200 filters (Pall) by applying gentle vacuum pressure.

After size fractionation, particles and zooplankton were broken up before plating (Fig. 1.5B). The zooplankton sample was homogenized using a tissue grinder (VWR Scientific) and vortexed for 20 minutes at low speed before concentration on 0.22 μm Supor-200 filters (Pall). These filters were then plated directly on selective media. Similarly, 5 μm and 1 μm filters were placed in 50 ml conical tubes with 50 ml sterile seawater and vortexed at low speed for 20 min to break up particles and detach bacteria from the filters. The supernatant was concentrated on 0.22 μm filters, and

both the original and supernatant filters were placed directly on media to collect isolates.

1.2.3 Strain isolation and identification

Isolates were obtained from TCBS plates (Accumedia or Difo) with 2% NaCl since this media has been shown to yield good correspondence in phylogenetic groups of vibrios detected by quantitative PCR and isolation [54]. After 2-3 days of growth, colonies were counted and re-streaked a total of three times alternately on Tryptic Soy Broth (TSB) (Difco) with 2% NaCl and media.

For classification of strains by sequencing, purified isolates were grown in marine TSB broth overnight; DNA was extracted using either a tissue DNA kit (Qiagen) or Lyse- N-Go (Pierce). Following the rationale of multilocus sequence analysis (MLSA), housekeeping genes were used for further strain characterization since these are unlikely to be under environmental selection. The partial *hsp60* gene sequence was amplified for all isolates as described previously [67]. For isolates with an *hsp60* sequence differing by more than 2% from an already characterized strain, the 16S rRNA gene was PCR amplified using primers 27F- 1492R and sequenced using the 27F primer [68]. The 16S sequence was used to identify the organism using the RDP classifier [69] and BLAST [70]. In cases where the *hsp60* gene either failed to amplify or the sequence diverged greatly from other vibrios, 16S rRNA gene sequencing confirmed that these isolates largely belonged to the genera *Pseudomonas*, *Shewanella*, *Pseudoalteromonas*, and *Agaravorans* (RDP Classifier) [69]. We excluded non-*Vibrionaceae* strains from further analysis.

To confirm relationships for *V. splendidus*, the most highly represented group among isolates, an additional gene (*mdh*) was sequenced. The partial *mdh* gene was amplified using primers *mdh*_mod.for (5'- GAY CTD AGY CAY ATC CCW AC -3') and *mdh*_mod.rev (5'- GCT TCW ACM ACY TCD GTR CCY G -3') (S. Preheim, unpublished data). Two additional housekeeping gene sequences were obtained (*pgi*, *adk*) for select groups of strains, using *pgi*.for (5 GAC CTW GGY CCW TAC ATG GT 3 - 3) and *pgi*.rev (5-CMG CRC CRT GGA AGT TGT TRT-3) (S. Preheim,

unpublished data) and *adk*.for (5- GTA TTC CAC AAA TYT CTA CTG G-3) and *adk*.rev (5- GCT TCT TTA CCG TAG TA- 3) [71]. All of these genes were amplified using the following PCR conditions: 2 min at 94°C followed by 32 cycles of 1 min each at 94°C, 46° and 72°C with a final step of 6 min at 72°C. Most genes were sequenced at least twice using forward and reverse primers. All sequencing was performed at the Bay Paul Center at the Marine Biological Laboratory, Woods Hole MA.

1.2.4 Phylogenetic tree construction and representation

The partial *hsp60*, *mdh*, *adk*, and *pgi* gene sequences yielded unambiguous alignments of 541, 422, 372, 395 nucleotides, respectively. Phylogenetic relationships were reconstructed using PhyML v.2.4.4 [72] with following parameter settings: DNA substitution was modeled using the HKY parameter [73]; the transition/transversion ratio was set to 4.0; PhyML estimated the proportion of invariable nucleotide sites; the gamma distribution parameter was set to 1.0; 4 gamma rate categories were used; a BIONJ tree was initially used; and, both tree topology and branch lengths were optimized by PhyML. Circular tree figures were drawn using the online iTOL software package [74]. To prevent numerical instabilities in AdaptMLs maximum likelihood computations, branches with zero length were assigned the minimal observed non-zero branch length: 0.001.

1.2.5 Empirical statistical testing

We employed empirical statistics to quantify evidence for differential environmental distribution of phylogenetic groups (Fig. 1.6). We first tested the overall association of phylogeny with our environmental data using a non-parametric parsimony-based metric. We assigned a different character to each of the environmental categories, and calculated the minimum number of character transitions needed to explain the data given the observed *hsp60* phylogeny. Although this test is likely to be overly conservative given the heterogeneous nature of our observed clusters, it nonetheless supported a highly significant correlation between phylogeny and both size fraction (p

$< 10^{-49}$) and season ($p < 10^{-79}$). Exact p-values were computed based on the algorithm of [55]. It was not possible to compute p-values for both season and size fraction together because the computational complexity of the algorithm grows exponentially with the number of character states.

We also employed non-parametric empirical statistics to test specific model predictions. We tested the hypothesis that each of the clusters identified by the model would be likely to arise by chance. To do this, we produced a 2×8 contingency table to test for any associations between cluster membership and distribution across environments. We used the Fisher exact test [75] as implemented in the R programming language to evaluate the significance of each association. The results are shown in Figure 1.4.

1.2.6 Overview of AdaptML

We developed a maximum likelihood method to help identify the boundaries of ecologically distinct populations and infer the ancestral habitat association of internal nodes in the strain phylogeny. The key to our method is a hidden variable mapping a 'projected habitat' to each node. We mathematically characterize each habitat as a discrete probability distribution, which describes the likelihood that a strain adapted to that habitat will be observed in each of our eight environmental categories. These distributions, which we refer to as emission probabilities in accordance with terminology used in machine learning, are not known a priori and must be learned from the data. Because of this probabilistic definition of habitats, a phylogenetic group spanning several environmental categories can still be considered an ecologically distinct population.

Using the habitat variables and isolate sequence-based phylogeny, we built a hidden-Markov model (HMM) [75] describing the evolution of habitat association (Figure 1.7). The probability that adjacent nodes in the phylogeny share the same habitat is a function of both the branch length separating them and a parameter that represents the rate at which a lineage can transition between habitats (the transition rate). The observed variables – the environmental category from which each strain

was sampled occur only at the leaves of the phylogeny. The parameters necessary for our model can be learned from the data according to the following algorithm:

1. **Initialize parameters:** We initialize 16 habitats, each with random emission probability distributions over the 8 environmental categories. The transition rate parameter is initialized to 10^{-1} transitions per substitution/site (relative to the gene phylogeny branch lengths).
2. **Infer the observed data likelihood, given phylogeny and parameter estimates:** We use a dynamic programming algorithm and the model parameters (transition rate and emission probabilities) to compute the likelihood of the observed data (environmental category for each isolate). Our computation proceeds in a manner identical to Felsenstein’s “pruning” method of computing likelihoods on a phylogenetic tree [76].
3. **Optimize parameters to maximize likelihood of observed data:** We estimate the probability that each internal node is associated with a given habitat by summing over all possible habitat assignments at other nodes (E-step). These probabilities are used (M-step) to update the:
 - (a) *Transition rate parameter:* We numerically optimize the transition rate to maximize the likelihood of the observed data.
 - (b) *Emission probability matrix:* We update the emission probability matrix by taking the matrix that maximizes the likelihood of the observed data given the marginal likelihoods for the habitat assignments at each of the phylogenys leaves.

We note that separating these two steps represents an approximation, as these two parameters are not strictly independent. The approximation, however, speeds up the implementation considerably as only one parameter (instead of $1 + 16 \times 7 = 113$) is optimized numerically.

4. **Test for convergence:** If the model parameters do not change significantly from the previous iteration, then the emission probabilities and the transition

rate are considered to have converged: continue to step (5). (A typical trajectory of emission probability convergence is shown in Figure 1.8. Note: the approximation identified in step 3 can lead to fluctuation near a likelihood maximum rather than actual convergence). Otherwise, return to step (2).

5. **Test for model complexity/redundancy.** If a pair of habitats has emission probability distributions that exhibit correlations greater than 0.90, they are merged into a single habitat and the algorithm continues from step (2). If no habitats are merged, the parameter estimation loop terminates. Although our approach employed manual inspection and empirical testing rather than a likelihood-based criterion for reducing model complexity [such as the AIC [77]], our algorithm can be easily extended to include a likelihood criterion. To test for overfitting, we performed simulations as described in the main text and Figure 1.10. We found that our scheme acts conservatively since it usually underestimated the true number of habitats. Figure 1.13 shows how the inferred habitats identified by the model vary with different cutoffs.

Once a set of model parameters has been learned, we utilize the following protocol to identify ecologically distinct and statistically significant populations.

1. **Infer node habitat assignments that maximize the joint probability of the observed data:** We rely upon a joint likelihood calculation to infer a single habitat assignment per ancestral node. To compute this likelihood, we use the parameter estimates inferred by the algorithm described above, which sums over all habitat assignments. Phylogenetic groups that share a common habitat association are taken as candidate ecological populations if they pass an empirical significance test.
2. **Empirical testing to identify ecologically distinct populations:** The assignment of nodes to habitats in step (1) identifies the most likely set of population boundaries, but may include some weakly predicted clusters. To filter low confidence ecological groupings, we estimate empirical p-values for

each clade and only report statistically significant ($p < 0.0001$) populations (Figure 1A & 1B). Empirical p-values are computed by comparing the likelihood of the parent node for a cluster to the likelihood observed at the same node in randomized trials where environmental assignments are shuffled, but phylogeny is maintained. For comparison, all possible clusters (with no significance cutoff) can be inferred from the full model results shown in Figure 1.11.

1.2.7 Detailed description of maximum likelihood model

Conditional likelihoods

The conditional likelihood describes the likelihood L that the leaves of a subtree exhibit their observed states, conditional on the subtree's root node k taking state s . This likelihood can be defined recursively, assuming two child nodes l and m

$$L_k(s) = \left(\sum_x P(x|s, t_l) L_{kl}(x) \right) \times \left(\sum_y P(y|s, t_m) L_{km}(y) \right) \quad (1.1)$$

where the function $P(x|s, t)$ represents the probability of transitioning between states x and s along some interval t . To reduce the number of fitted parameters early in our model, we use the simplifying assumption that all state transitions can be described using the same transition rate parameter μ . Thus, we compute the probability of transitions between states as

$$P(x|y, t) = \frac{1}{h} (1 - e^{-h\mu t}) \quad (1.2)$$

and the probability of remaining in the same state

$$P(x|x, t) = \frac{1}{h} (1 + (h - 1) \times e^{-h\mu t}) \quad (1.3)$$

which is analogous to the Jukes-Cantor model for nucleotide sequence evolution [78]. Note that if k is a leaf node in state s with observed environment o , the likelihood is drawn from the emission probability matrix P_e :

$$L_k(s) = P_e(o|s) \quad (1.4)$$

Marginal and joint likelihoods

To compute the marginal likelihood that a node k has state s , we combine the conditional likelihoods:

$$ML_k(s) = \frac{1}{K} \times \prod_t \left[\sum_x P(x|s, t_{kl}) L_l(x) \right] \quad (1.5)$$

where the l are drawn from the set of nodes adjacent to k , and K is a normalization factor such that

$$\sum_s ML_k(s) = 1 \quad (1.6)$$

To compute the likelihood of the observed data for the single best assignment of habitats to nodes (17), the summation terms in the likelihood formula are replaced with **max** operations. This is equivalent to maximizing the “joint” likelihood:

$$L_k(s) = \left(\max_x P(x|s, t_{kl}) L_l(x) \right) \times \left(\max_y P(y|s, t_{km}) L_m(y) \right) \quad (1.7)$$

The backtracking process used to keep track of the max arguments is analogous to the Viterbi algorithm for finding the most probable state path in an HMM (16).

Parameter estimation

As probability distributions, each set of emission probabilities must satisfy

$$\sum_o P_e(o|s) = 1 \quad (1.8)$$

Because leaf nodes are independent samples of their emission probability distributions (given their state assignment), we use a weighted-average approach to calculating the emission probability matrix P_e

$$P_e(o|s) = \frac{1}{K} \times \sum_k ML_k(s) \times \Delta(o, o_k) \quad (1.9)$$

where $ML_k(s)$ is the marginal likelihood that node k is adapted to habitat s , the function $\Delta(x, y)$ equals 1 if and only if x equals y (and is 0 otherwise), and K is a normalization factor.

We learn the transition rate parameter μ by numerical optimization to maximize the likelihood of the observed data (summed over all values for the hidden variables).

1.2.8 Defining ecologically coherent and significant clusters

We use empirical testing to estimate the statistical significance associated with the likelihood value computed for each node in the maximum (joint) likelihood assignment of habitats to nodes (given the final parameter estimates). Each trial preserved the phylogeny, the inferred habitat assignments, and the habitat and transition rate parameters, but environmental categories at the leaves were shuffled. The maximum 'joint' likelihoods at each node were compiled over all trials and used as empirical background probability distributions.

We use empirical testing to estimate the statistical significance associated with the likelihood value computed for each node in the maximum (joint) likelihood assignment of habitats to nodes (given the final parameter estimates). Each trial preserved the phylogeny, the inferred habitat assignments, the habitat and transition rate parameters, and the frequency of the various environmental categories; however, each leaf was randomly assigned an environmental category in each trial. The maximum joint likelihoods at each node were compiled over all trials and used as empirical background probability distributions.

Using the habitat associations learned from the original (non-randomized) data set, we identified internal nodes where habitat transitions were inferred to take place. These nodes were then iterated through using a post-fix traversal:

Nested clusters

At each of these transitional nodes, a second, pre-fix traversal took place. If the subtree rooted by the current node possessed a likelihood greater than that observed in 99.99% of the random trials and 90% of its leaves shared the current nodes habitat assignment, we recognized the subtree as an ecologically coherent and statistically significant cluster. This latter cutoff (90% coherence) was necessary to ensure meaningful clusters because the likelihood at a parent node can be unusually high when two nearly significant (but non-identical) child groups are combined. Requiring a majority of child nodes to have the same predicted model state as the parent has the effect of identifying clusters that correspond more directly to single putative populations. Making the threshold too high (e.g., 100%) would eliminate some larger clusters that have a small, internal nested cluster.

To enable the discovery of nested clusters, identified clusters were pruned from the phylogeny. Nodes representing clades that contained the cluster had their joint-likelihoods modified so that they no longer incorporated information from the pruned groups. If the clade rooted by the current node did not satisfy both the p-value and 90% coherence thresholds, the second recursion would descend to the current nodes children. The second recursion only terminated if the current node was either a leaf, itself a habitat transition node, or determined to root an ecologically coherent and statistically significant cluster.

Supplementary Figures

| | Temperature | Chlorophyll a ¹ | DOC ² | TDN ² | NO ₃ ⁻ +NO ₂ ⁻ | NH ₄ ⁺ | TDP ² | PO ₄ ³⁻ |
|---------------------|-------------|----------------------------|------------------|------------------|--|------------------------------|------------------|-------------------------------|
| | [°C] | [µg/L] | [mg C/L] | [mg N/L] | [µg N/L] | [µg N/L] | [µg P/L] | [µg P/L] |
| Spring (4/28/06) | 11 | 4.07 | 2.11 | 0.17 | 9 | 189 | 18 | 14 |
| Fall (9/6/06) | 16 | 6.03 | 2.28 | 0.27 | 5 | 144 | 24 | 25 |

¹ measured using overnight extraction in 90% acetone (19)

² DOC = dissolved organic carbon, TDN = Total Dissolved Nitrogen, TDP = total dissolved phosphorous.
All chemical analyses were performed at the University of New Hampshire, Durham, NH.

Figure 1.3: Temperature and nutrient concentrations on sampling dates.

| Group | P-value |
|-------|-----------|
| 1 | 1.16 E-07 |
| 2 | 1.04 E-16 |
| 3 | 5.06 E-15 |
| 4 | 1.70 E-01 |
| 5 | 1.11 E-02 |
| 6 | 1.26 E-04 |
| 7 | 7.54 E-10 |
| 8 | 2.46 E-05 |
| 9 | 2.35 E-07 |
| 10 | 7.19 E-07 |
| 11 | 1.26 E-13 |
| 12 | 2.62 E-03 |
| 13 | 5.13 E-10 |
| 14 | 1.16 E-06 |
| 15 | 1.01 E-02 |
| 16 | 1.13 E-08 |
| 17 | 4.03 E-07 |
| 18 | 2.81 E-03 |
| 19 | 1.08 E-08 |
| 20 | 1.80 E-06 |
| 21 | 2.31 E-13 |
| 22 | 7.02 E-03 |
| 23 | 3.32 E-06 |
| 24 | 6.06 E-14 |
| 25 | 2.76 E-03 |

Figure 1.4: Empirical significance testing of ecologically differentiated clusters predicted by the model. Fishers exact test was used to identify significant associations between clusters and environments as described in the SOM. Group numbers correspond to Figure 1.1 in the main text.

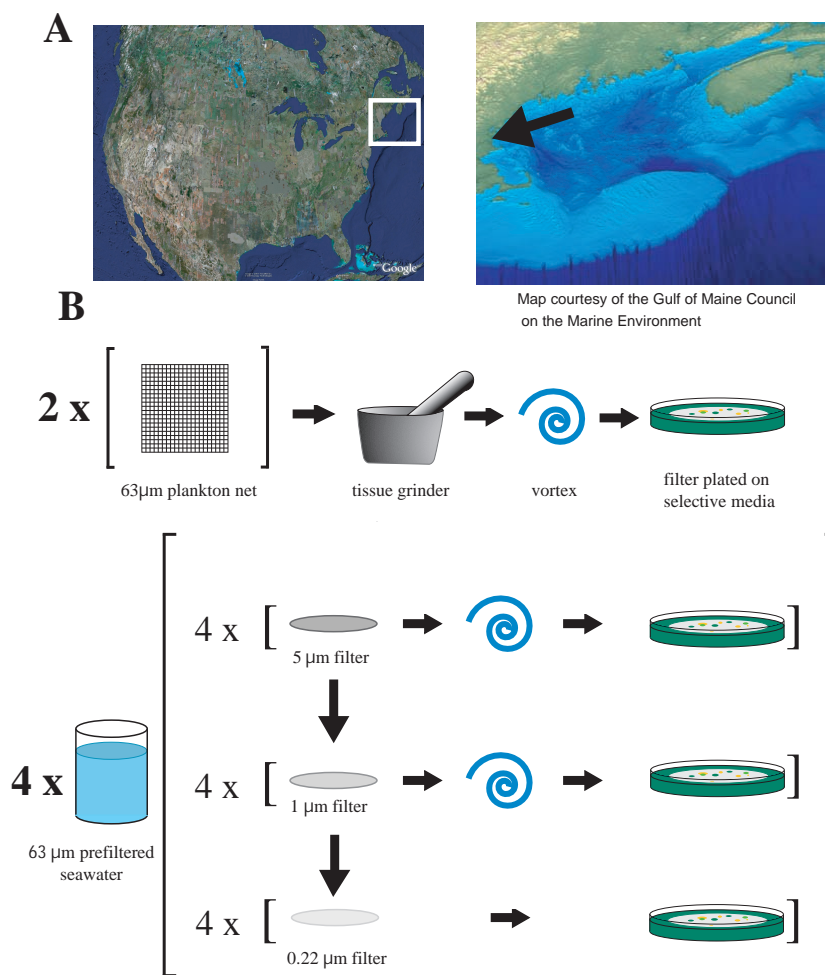


Figure 1.5: Sampling site and outline of sampling strategy for determination of bacterial distribution among seasons and size fractions in coastal water. **(A)** Sampling location shown on a map of North America (left) with a white box depicting the bounds of the picture at right: the Gulf of Maine. The arrow points to the sampling location, Plum Island Sound, MA. **(B)** Protocol for obtaining size fractionated of bacterial seawater isolates using sequential filtration and plating on selective media.

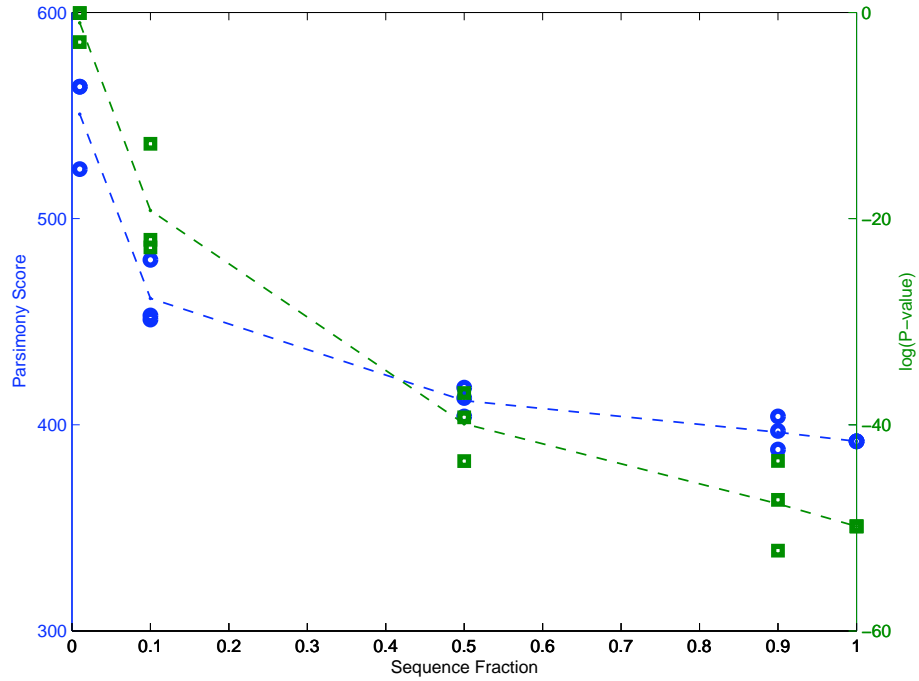


Figure 1.6: Empirical statistical testing for ecological association of phylogenetic clades. The likelihood that the observed parsimony score (or lower) for size fraction data might have arisen by chance was calculated [using the method of [55]] for a series of trees, inferred using subsets of strains from the *hsp60* gene alignment. To test the effect of statistical uncertainty on the inferred association, 1%, 10%, 50%, 90%, or 100% of the sequence data was randomly selected and used to construct the phylogeny (with 3 replicates each). The parsimony score for each such tree is shown with green dots corresponding to the left axis; the p-value of obtaining that parsimony score by chance is shown with blue dots corresponding to the right axis. These results are based on size fraction data only, but similar results are obtained with season data.

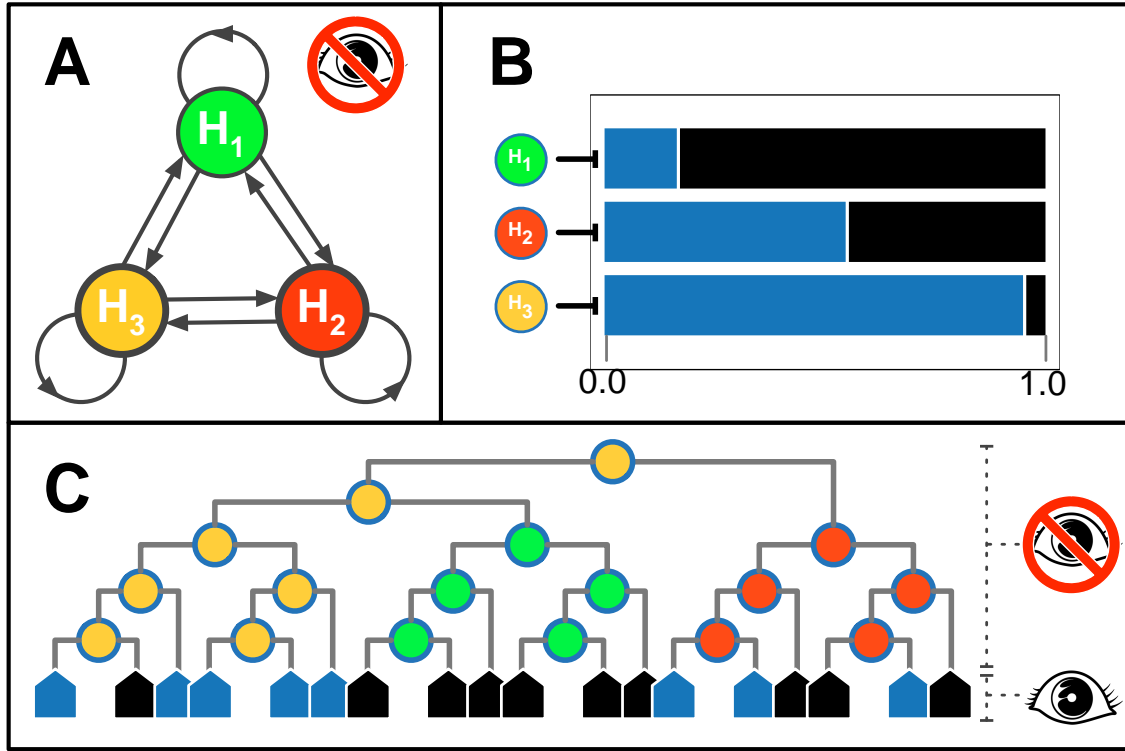


Figure 1.7: Overview of hidden Markov model components. **(A)** Habitats represent hidden (latent) variables in the model; experiments do not directly measure what habitat a strain is adapted to. Two adjacent nodes on the phylogeny may differ in their habitat assignment according to the rate of habitat transition (arrows between habitats). In our simple model, all transition rates are equivalent. **(B)** Associated with each habitat is an emission probability distribution describing how likely a strain associated with a particular habitat (H_1 - H_3) is sampled from a given environment (blue or black). Bars in this cartoon depict hypothetical probability distributions; strains adapted to habitat 1 have higher probability of being observed in the black environment than in the blue environment. **(C)** Our model maps a habitat onto each node in the phylogeny such that it maximizes the probability of the observed data (at the leaves). Observations are limited to the leaves of the phylogeny, as we cannot directly sample ancestral strains. In the example shown, the mostly blue group is mapped to habitat 3, the mostly black group to habitat 1, and the heterogeneous group is mapped to habitat 2.

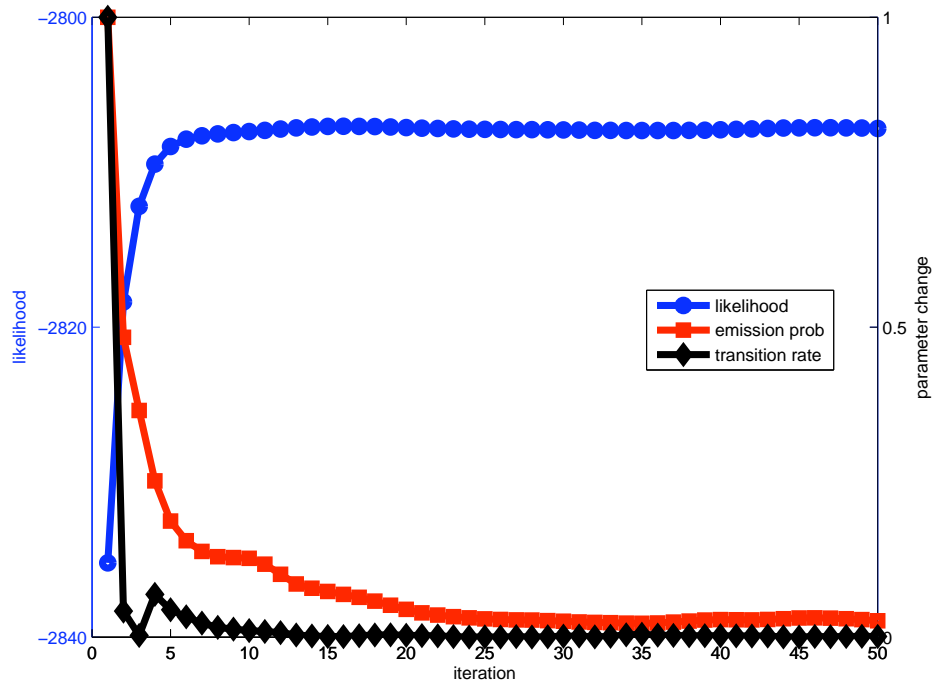


Figure 1.8: Convergence of iterative parameter optimization. During the parameter optimization, both the average change in probability for components of the emission probability matrix (red line) and the change in transition rate (black line) decrease rapidly. The overall log-likelihood of the observed data converges in concert with the parameter estimates (blue line).

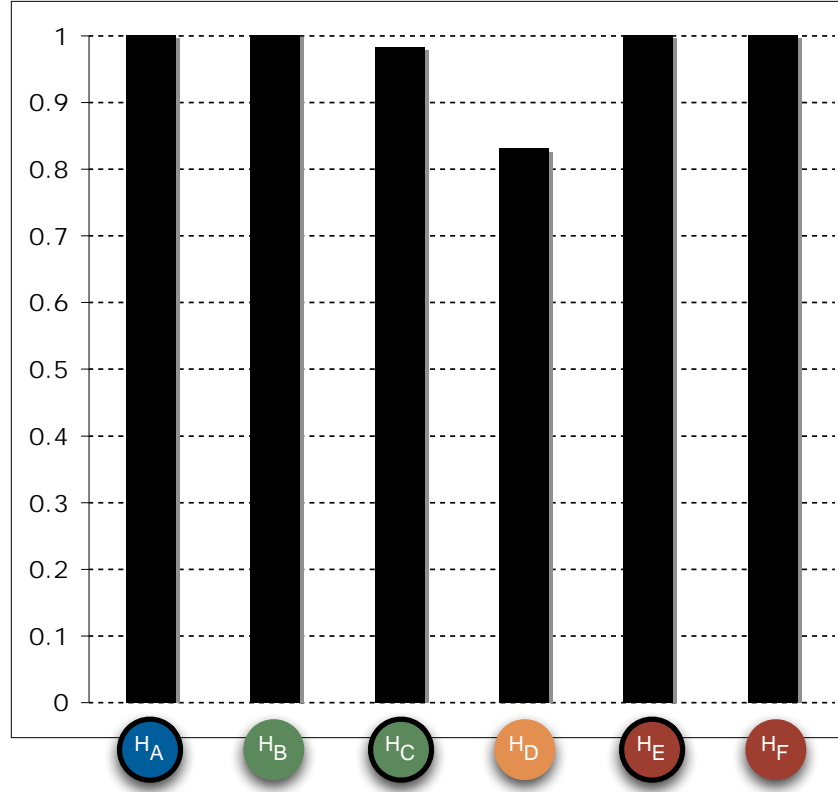


Figure 1.9: Reproducibility of inferred habitats. Sixty independent trials of the iterative habitat learning process were performed. Shown are the frequencies of occurrence of the habitats presented in Figure 1.1. The habitat similarity cut-off for counting a match was an average emission probability difference < 0.10 over all environmental categories. As expected, the least abundant habitat in our data, H_D (see Fig. 1.8) is the least reproducible, although even this habitat is identified in 83% of trials.

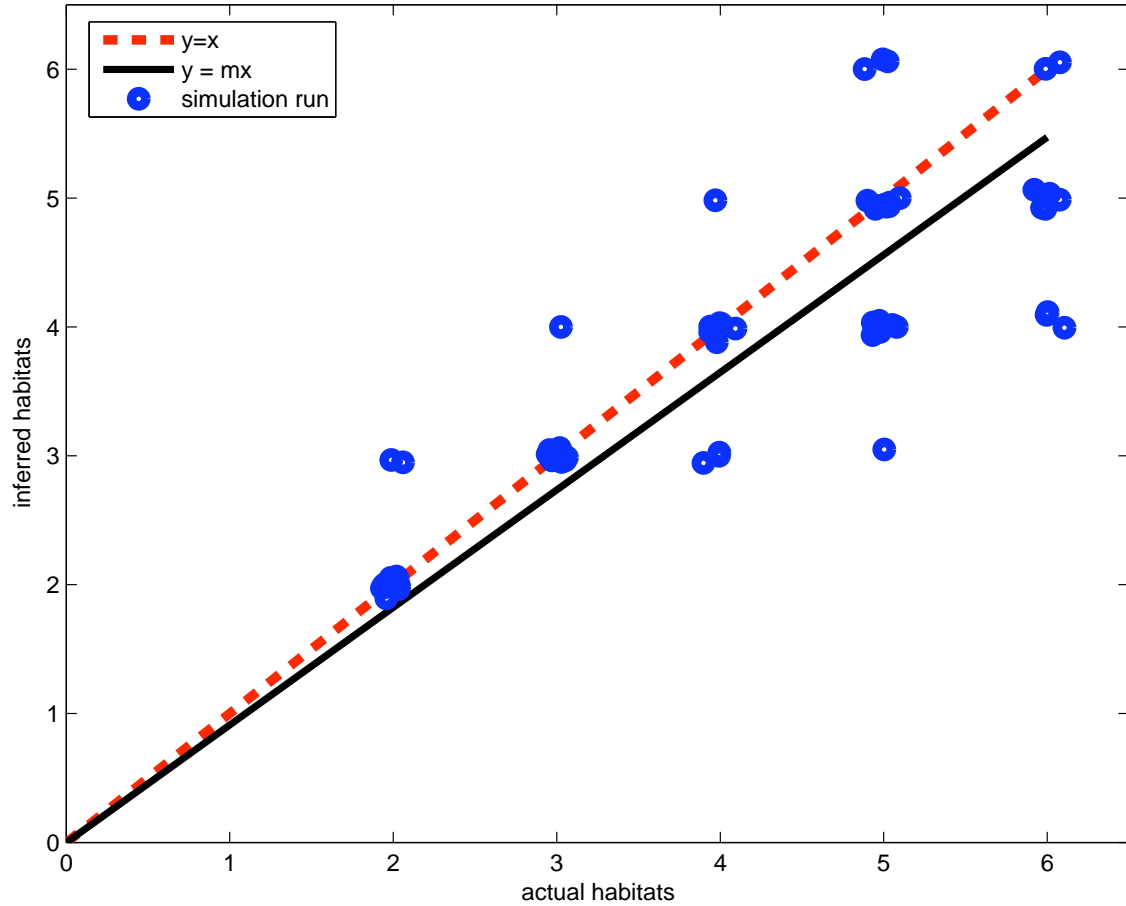


Figure 1.10: Number of habitats inferred from simulated data. We generated 120 simulated datasets and compared the number of inferred habitats to the true number. Each dataset was randomly assigned between 2 and 6 habitats; these habitats were distributed over environmental categories by randomly partitioning the interval (0,1) according to a uniform distribution, but requiring that each successive partitioning must occur at a larger value than the last (random partitioning without this restriction leads to a large number of similar generalist habitats). These habitats were mapped onto the set of clusters learned from the *Vibrionaceae* data (Figure 1A). Shown are the number of habitats inferred for these trials versus the number actually present (a small amount of Gaussian noise was added to each data point so that the data points could be discerned). These results suggest that the habitat inference algorithm is generally conservative, inferring fewer rather than more than the true number of habitats.

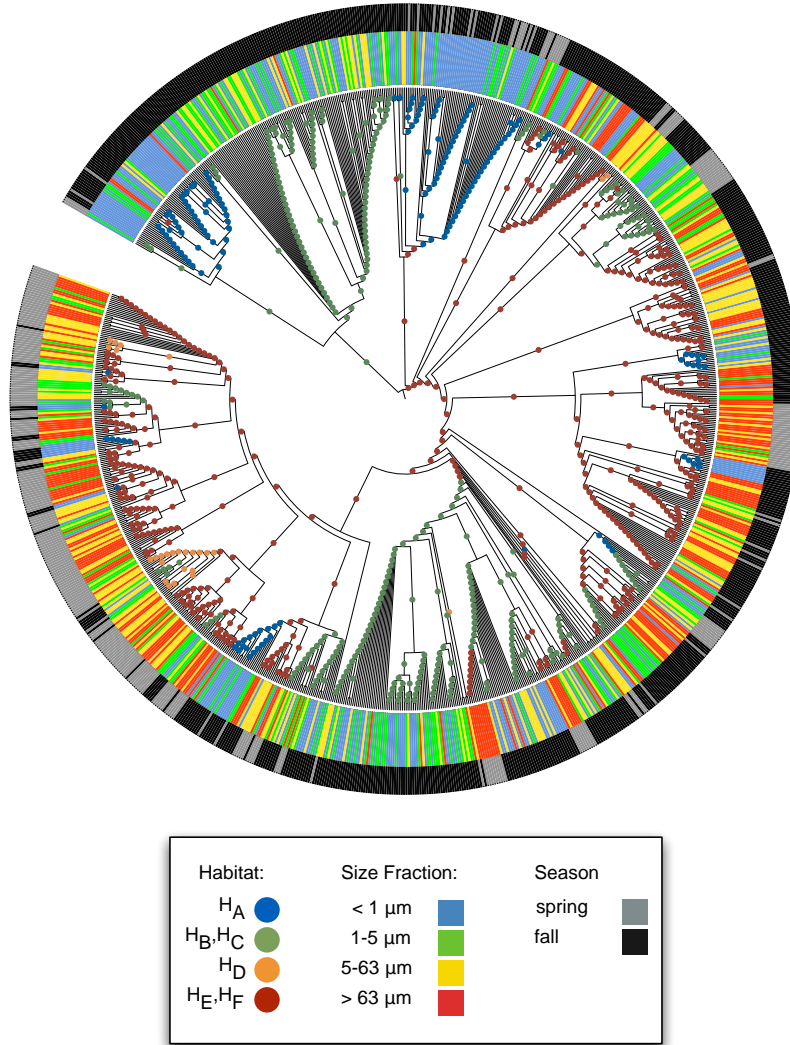


Figure 1.11: Inferred habitat associations for all ancestors of sequenced *Vibrio* strains. The rings surrounding the tree represent the season (outer) and size fraction (inner) from which strains were isolated. The maximum likelihood assignment of nodes to habitats is shown for all nodes, regardless of the confidence of each prediction (only confident assignments are shown in Figure 1.1A). Colored circles on each branch indicate the habitat assignment (H_A - H_F , as in Figure 1.1A) for the node immediately below that branch (see above legend for color scheme). Branch lengths are adjusted to aid visualization and do not represent evolutionary distances.

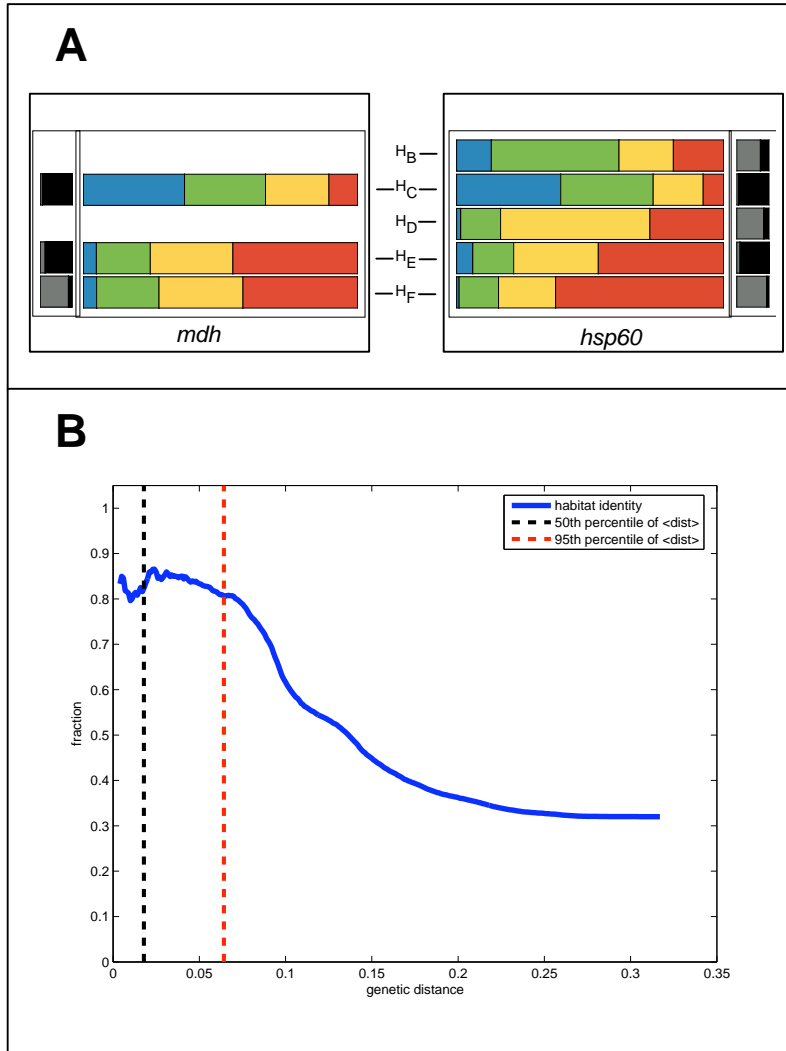


Figure 1.12: Comparison of habitat inference on different gene phylogenies (*hsp60* and *mdh*) for *Vibrio splendidus* strains. **(A)** Juxtaposition of habitats learned from the *mdh* and *hsp60* datasets for *V. splendidus* strains only; habitats are labeled to allow comparison with habitats predicted for all *Vibrionaceae* in Fig. 1.1. Emission probabilities are normalized by the total number of isolates obtained in each environmental sample to reduce the effects of sampling bias. As expected, fewer habitats are identified from the *mdh* phylogeny, which has a lower rate of nucleotide divergence and thus is less well-resolved. **(B)** Comparison of habitat assignments to nodes. Because it is difficult to map the internal nodes between topologically distinct trees, the habitat assignment for the last common ancestor of each pair of *V. splendidus* strains was compared. If both corresponded to the same habitat (H_C , H_E , or H_F in both phylogenies), they were considered to be in agreement, otherwise they were considered to be in disagreement. The fraction of nodes in agreement is shown as a function of increasing genetic distance between the pairs of strains considered (in the *hsp60* phylogeny). The black and red lines indicate distances that include 50% and 95% of strains within the same cluster in main text Figure 1.1, respectively.

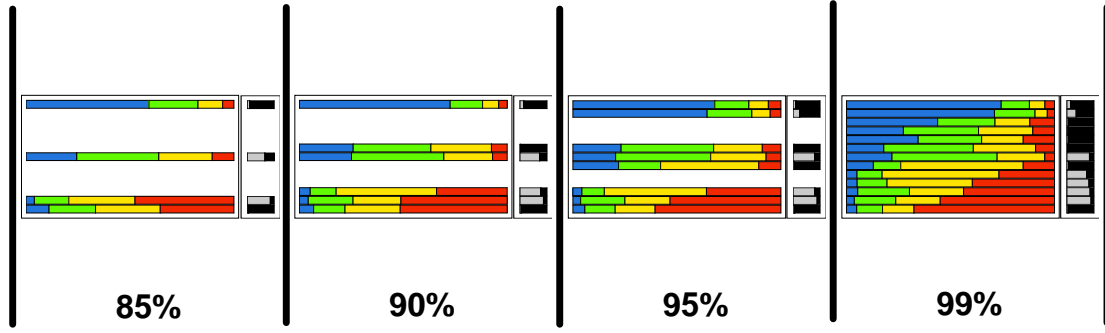


Figure 1.13: Influence of the model complexity/redundancy parameter on inferred habitats. Clusters are merged during the model fitting procedure when the vectors describing their distribution across environments are more than 90% correlated. As this cutoff is varied, slightly different habitats are observed. At 85%, habitat H_C is not recovered; at higher values additional habitats become more redundant suggesting that the 90% cutoff allows conservative recovery of characteristic projected habitats.

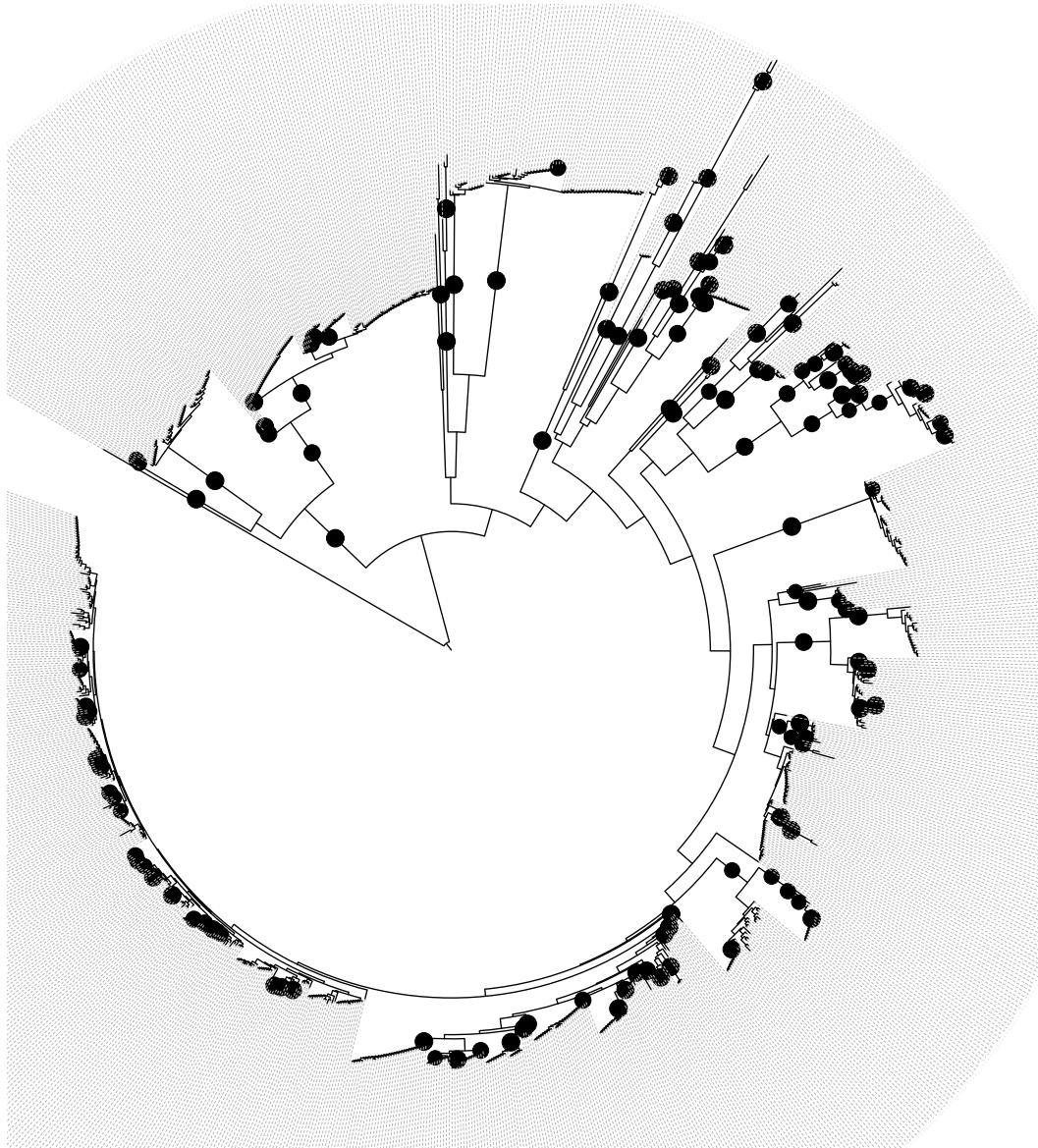


Figure 1.14: Statistical uncertainty in the *hsp60* gene tree by constructing 100 trees using non-parametric bootstrap re-sampling from the *hsp60* alignment. Clades supported in greater than 80% of bootstraps are indicated with a black dot.

Chapter 2

Rapid evolutionary innovation during an Archean Genetic Expansion

Lawrence A. David, Eric J. Alm

This chapter is presented as it was submitted for publication. Corresponding
Supplementary Material is appended.

Chapter 2

Rapid evolutionary innovation during an Archean Genetic Expansion

A natural history of Precambrian life remains elusive because of the rarity of microbial fossils and biomarkers [79, 80]. The composition of modern day genomes, however, may bear imprints of ancient biogeochemical events [81–83]. We have employed an explicit model of macroevolution including gene birth, transfer, duplication and loss to map the evolutionary history of 3,968 gene families across the three domains of life. We observe that horizontal gene transfer (HGT) is the primary source of new genes in prokaryotes, while duplication dominates in eukaryotes. Inter-domain gene transfer is rare compared to intra-domain transfer with the notable exception of massive Bacteria-Eukarya transfer events that correspond to the endosymbiosis of the mitochondria and chloroplasts [84, 85]. Surprisingly, we find that a brief period of genetic innovation during the Archean eon gave rise to 27% of major modern gene families. Genes born during this period are especially likely to be involved in electron transport, while later genes exhibit a gradually increasing usage of molecular oxygen. Our results demonstrate that reconstructing the complex interplay between organismal and geochemical evolution over Earth history is becoming a tractable goal.

Introduction

Describing the emergence of life on our planet is one of the grand challenges of the Biological and Earth sciences. Yet the roughly three-billion-year history of life preceding the emergence of hard-shelled metazoans remains obscure [79]. To date, the best understood event in early Earth history is the Great Oxidation Event, which is believed to follow the invention of oxygenic photosynthesis by the ancestors of modern cyanobacteria [86] (though the precise timeline remains controversial [80]). If DNA sequences from extant organisms bear an imprint of this event, then we can use them to make and test predictions [81–83]; e.g., genes that use molecular oxygen will be confined to a group of organisms emerging after the Great Oxidation Event. Transfer of genes across species, however, can obscure patterns of descent and disrupt our ability to correlate gene histories with the geochemical record [87]. Widely distributed genes, for example, may descend from a Last Universal Common Ancestor (LUCA) as widely believed to be the case for the translational machinery [88], or may have been dispersed by HGT [12, 89], as in the case of antibiotic resistance cassettes.

Methods Overview

We developed a new phylogenomic method, AnGST (Analyzer of Gene and Species Trees), to account for the confounding effects of HGT by comparing individual gene phylogenies to the phylogeny of organisms (the Tree of Life). We refer to this process as tree reconciliation and provide a detailed description of the AnGST algorithm in the Supplementary Information. Unlike some previous methods [24, 25, 27], AnGST uses the topology of the gene family tree rather than just its presence/absence across genomes and can infer duplication, HGT, and loss events. Importantly, AnGST also accounts for uncertainty in gene trees by incorporating reconciliation into the tree-building process: the tree that minimizes the evolutionary cost function, but is still supported by the sequence data, is chosen as the best gene tree. Simulated trees inferred with this method are more accurate than trees based on a maximum likelihood model of sequence data alone (Figure 2.5). Thus, tree building methods such as

AnGST that explicitly model macroevolutionary events may have utility in phylogenetic inference [33]. We used a previously described Tree of Life [90] to reconcile gene families, although we note that our key results were consistent when using 30 alternative reference trees, including those that used the Archaea or Eukarya as outgroups (Figs. 2.11, 2.12). Ensuring proper causality in a large reconciliation (i.e., avoiding the “grandfather paradox” in which a gene is inferred to be its own ancestor) is a computationally intractable problem in general [31], which we overcome by explicitly modeling the timing of evolutionary events based on a chronogram constructed from our reference tree. A conservative set of eight temporal constraints was selected from the geochemical and paleontological literature (Table 2.1), and the PhyloBayes software package was used to infer a range of divergence times for each ancestral lineage on the reference tree [91]. We did not apply temporal constraints to lineage ages on the gene trees.

Results

Domain-specific Macroevolutionary Trends

For 3,968 extant gene families [106], AnGST predicted a total of 109,452 speciation, 38,575 HGT, 14,021 gene duplication, and 35,252 gene loss events (Figure 2.1A). The abundance of HGT events (on average 9.7 per gene family) underscores the evolutionary importance of gene transfer in prokaryotic genome structure (Figure 2.15). Domain-specific preferences in the types of macroevolutionary events emerge in Figure 2.1B. On a per-gene basis, gene transfer is 2.1 times more likely in bacteria than in eukaryotes, while duplications are 4.4 times more likely in eukaryotes than in bacteria. The rate of HGT in eukaryotes is likely to be an overestimate because we did not consider eukaryote-only gene families. The bias toward duplication in eukaryotes is consistent with known domain-specific traits, such as unequal crossing-over, whole-genome duplication events, and reduced selection against large genome sizes [16, 107, 108]. Interdomain transfers comprise a minority (16.1%) of HGT events but exhibit significant over-representation of HGT from alpha-proteobacteria to an-

| | Event | Constraint | Evidence |
|---|--|----------------------------------|---|
| 1 | Last universal common ancestor arises | < 3850 Ma | Carbon isotope fractionation [92, 93] |
| 2 | Cyanobacteria emerge | > 2500 Ma | Traces of an aerobic nitrogen cycle [94], changes in redox-metal enrichments [95], and sulfur isotope fractionation data [96, 97] indicate oxygenic photosynthesis; traces of 2 α -methylhopane biomarkers [98] indicate cyanobacterial presence |
| 3 | Eukaryotes diverge from Archaea | > 2670 Ma | Preserved sterane biomarkers [98, 99] |
| 4 | Akinetes diverge from cyanobacteria lacking cell differentiation | > 1500 Ma | Akinete microfossils [100] |
| 5 | Archaeplastida emerge | > 1198 Ma | Red algae microfossils [101] |
| 6 | Animals emerge | > 635 Ma | Preserved demosponge steranes [102] |
| 7 | Tetrapods emerge | (a) < 385 Ma (b) > 359 Ma | (a) Tetrapod precursor dating [103]; (b) Tetrapod fossil dating [104] |
| 8 | <i>Buchnera</i> diverge from <i>Wigglesworthia</i> | > 160 Ma | Fossil history of <i>Buchnera</i> 's aphid hosts [105] |

Table 2.1: Temporal constraints used to construct chronogram. Eight temporal constraints that could be directly linked to fossil or geochemical evidence were used to estimate divergence times on the Tree of Life (Figure 2.10).

cient eukaryotes ($p=3.3 \times 10^{-7}$ Wilcoxon rank sum test) and from cyanobacteria to plants ($p=8.3 \times 10^{-6}$ Wilcoxon rank sum test). These results likely reflect the ancient endosymbioses that gave rise to the mitochondrial and chloroplast organelles [84, 85]. Functional analysis of HGT from the alpha-proteobacteria to the ancestral eukaryotes reveals significant enrichment for energy metabolism genes ($p=1.6 \times 10^{-6}$ Fisher’s Exact Test), further supporting an association between these HGT and an energy-producing endosymbiosis (Figure 2.18). HGT from cyanobacteria to *Arabidopsis thaliana* are also enriched for energy-producing genes ($p=3.9 \times 10^{-3}$ Fisher’s Exact Test), as well as translation-related genes ($p=4.4 \times 10^{-5}$ Fisher’s Exact Test) which likely reflect the migration of 70S ribosomal proteins from the chloroplast to the plant nucleus [109].

An Archean Genetic Expansion

Gene histories reveal dramatic changes in the rates of gene birth, duplication, loss, and HGT over geologic time scales (Figure 2.2). The most striking feature of the overall gene flux depicted in Figure 2 is a burst of de novo gene family birth between 3.33-2.85 Ga which we refer to as the Archean Genetic Expansion (AGE). This window gave rise to 26.8% of extant gene families and coincides with a rapid bacterial cladogenesis. A spike in the rate of gene loss (~ 3.1 Ga) follows the AGE and may represent consolidation of newly evolved phenotypes, as ancestral genomes became specialized for newly emerging niches. After 2.85 Ga, the rates of both gene loss and gene transfer stabilize at roughly modern-day levels. The rates of de novo gene birth and duplication after the AGE appear to show opposite trends: de novo gene family birth rates decrease and duplication rates increase over time. The near absence of de novo birth in modern times likely reflects the fact that ORFan gene families, which are widespread across all major prokaryotic groups, are not considered in this study [110]. The excess of gene duplications and ORFans in modern genomes suggests that novel genes from both sources experience high turnover and rarely persist over long evolutionary time scales.

What evolutionary factors were responsible for the period of innovation marked by

the AGE? While we cannot provide an unequivocal answer to this question using gene birth dates alone, we can ask whether the functions of genes born during this time suggest plausible hypotheses. In general, birth of metabolic genes is enriched during the AGE, especially those involved in energy production and coenzyme metabolism (Table 2.17), but further inspection also reveals an enrichment for metabolic gene family birth prior to the AGE. To focus on specific metabolic changes linked to the AGE we: (i) grouped genes according to the metabolites they used; and (ii) we directly compared the occurrence of these metabolites in genes born during the AGE to their abundance prior to the AGE. The results are striking: the AGE-specific metabolites (positive bars, Figure 2.2 inset) include most of the compounds annotated as redox/ e^- transfer (blue bars), with Fe-S-, Fe-, and O_2 -binding gene families showing the most significant enrichment (False Discovery Rate < 5%, Fisher’s exact test). Gene families that use ubiquinone and FAD (key metabolites in respiration pathways) are also enriched, albeit at slightly lower significance levels (False Discovery Rate < 10%). The ubiquitous NADH and NADPH are a notable exception to this trend and appear to have played a role early in life history. By contrast, enzymes linked to nucleotides (green bars) exhibited strong enrichment in genes of more ancient origin than the AGE.

The observed metabolite usage bias suggests that the AGE was associated with an expansion in microbial respiratory and electron transport capabilities. Proving this association to be causal is beyond the power of our phylogenomic model. Yet this hypothesis is appealing because more efficient energy conservation pathways could increase the total free energy budget available to the biosphere, possibly enabling the support of more complex ecosystems and a concomitant expansion of species and genetic diversity. We note, however, that while the use of oxygen as a terminal electron acceptor would have significantly increased biological energy budgets, oxygen-utilizing genes are only enriched toward the end of the AGE (Figure 2.14). Thus, the earliest redox genes we identified as part of the AGE were likely to be used in anaerobic respiration or oxygenic/anoxigenic photosynthesis, although some may have been co-opted later for use in aerobic respiration pathways.

Phylogenomic evidence for ancient changes in global redox potential

Our metabolic analysis supports an increasingly oxygenated biosphere following the AGE, as the fraction of proteins utilizing oxygen gradually increases from the AGE until the present day (Figure 2.3; $p=3.4 \times 10^{-8}$, two-sided Kolmogorov-Smirnov test). Further indirect evidence of rising oxygen levels comes from compounds that are sensitive to global redox potential. We observe significant increases over time in the usage of the transition metals copper and molybdenum (Figure 2.3; False Discovery Rate $< 5\%$, two-sided Kolmogorov-Smirnov test), which is in agreement with geochemical models of these metals' solubility in increasingly oxidizing oceans [82, 83] and with the growth of molybdenum enrichments from black shales that suggests molybdenum began accumulating in the oceans only after the Archean eon [111]. Our prediction of a significant increase in nickel utilization accords with geochemical modeling predictions of a 10X increase in dissolved nickel concentration between the Proterozoic and modern day [82], but conflicts with a recent analysis of banded iron formations that inferred monotonically decreasing maximum deposited nickel concentrations from the Archean onwards [112]. The abundance of enzymes using oxidized forms of nitrogen (N_2O and NO_3) also grows significantly over time, with 1/3 of nitrate-binding gene families appearing at the beginning of the AGE and 3/4 of nitrous oxide-binding gene families appearing by the AGEs end. The timing of these gene family births provides phylogenomic evidence for an aerobic nitrogen cycle by the Late Archean [94].

One striking discrepancy between our phylogenomic patterns and geochemical predictions, however, is a modest but significant increase in iron-using genes over time (Figure 2.3; False Discovery Rate $< 5\%$, two-sided Kolmogorov-Smirnov test). The cessation of iron formation deposition roughly 1.8 Ga and ocean chemistry models indicate that iron usage should decrease following the Archean, as declining iron solubility in oxygenated ocean surface waters and sulfide-mediated iron removal from anoxic deeper waters combined to reduce overall iron bioavailability [113]. The counterintuitive phylogenomic prediction may reflect the confounding effect of evolutionary inertia, whereby microbes in the face of declining iron availability could have found

more success evolving a handful of metal-acquisition proteins (e.g. siderophores), rather than replacing a host of iron-binding proteins. Alternatively, the insolubility of iron in modern oceans may be offset by large existing organic pools of reduced iron.

A precise timeline for oxygen availability is beyond the resolution of our relaxed molecular clock approach and remains a contentious topic in organic geochemistry [80, 98, 99]. Nonetheless, our results suggest an Archean biosphere containing some of the basic components required for oxygenic photosynthesis and respiration, despite the fact that appreciable oxygen levels do not appear in the geological record until much later (roughly 2.5 Ga) [114, 115]. Although our results are consistent with recent biomarker-based evidence for early oxygenesis [99], special caution should be used in comparing the molecular and geological dates. Divergence times for deep nodes on our reference chronogram are uncertain (Figure 2.16), and they are partially based on the constraint that the LUCA could be as old as the earliest evidence for life (3.85 Ga) [92], even though the LUCA is likely a descendant of the first life form. Furthermore, although the PhyloBayes dates include uncertainty estimates that are accurate given the assumptions of the CIR model [91], an alternative, semi-parametric approach implemented in r8s [116] results in a much younger date of 2.75-2.5 Ga for the AGE (compared to 3.33-2.85 Ga for PhyloBayes) which is closer to the Great Oxidation Event (Figure 2.12). Here we present mainly results from the PhyloBayes program, because it allowed us to explicitly account for uncertainty in the timing of inferred events. With such disparate phylogenetic estimates of the timing of the most ancient lineages, a chronology for the AGE and the evolution of oxygen-producing genes will require a careful integration of both geochemical and genetic data.

Implications

Using just eight temporal constraints as our geochemical and paleontological guides, we have shown that whole genome sequence data can be used to infer details of microbial evolution during the Archean eon and can recount global changes in redox-sensitive compound bioavailability following the evolution of oxygenesis. Still, our phylogenomic approach represents only a first step toward linking whole genome

sequence data to early Earth history. By connecting events in gene histories to events in Earth history, hypotheses of enzyme or pathway presence/absence can be used to make testable predictions about when metabolic signatures should first appear in the geochemical record. Conversely, geochemical hypotheses may be tested against predictions of extant metabolisms (as we demonstrate using the rise of oxygen). This may admit useful new lines of evidence for geochemical theories that suffer from gaps in the rock record. Successive refinement of phylogenomic models against geochemical constraints may eventually yield an abundant and reliable source of Precambrian fossils: modern-day genomes.

2.1 Figures

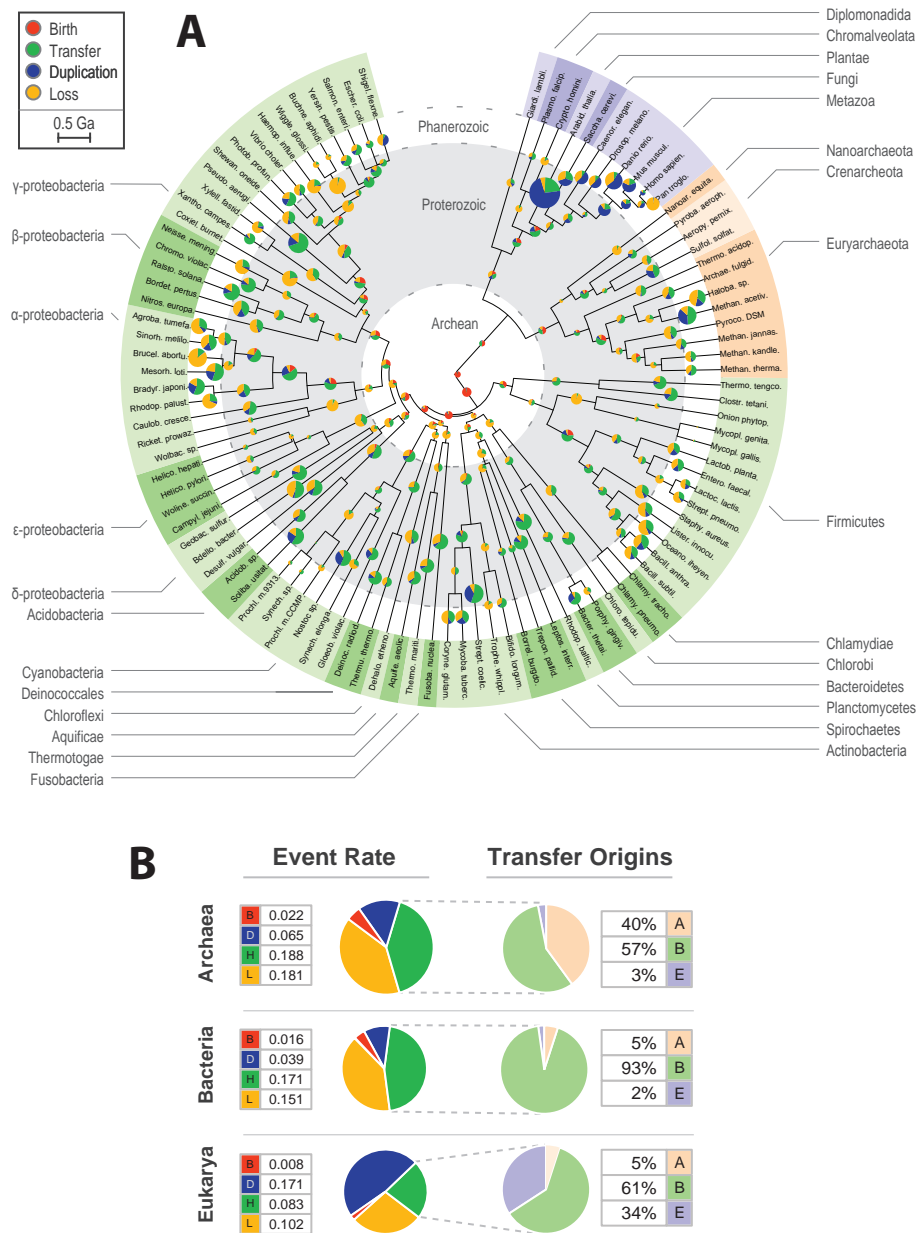


Figure 2.1: Evolutionary events by lineage. (A) The number of macroevolutionary events is mapped to each lineage on an ultrametric Tree of Life and visualized using the iTOL website [74]. Pie chart area denotes the number of events, and color indicates event type: gene birth (red), duplication (blue), HGT (green), and loss (yellow). (B) The average number of events per gene copy is separated by domain and the origins of HGT events are depicted: Bacteria (green), Archaea (beige), Eukarya (violet).

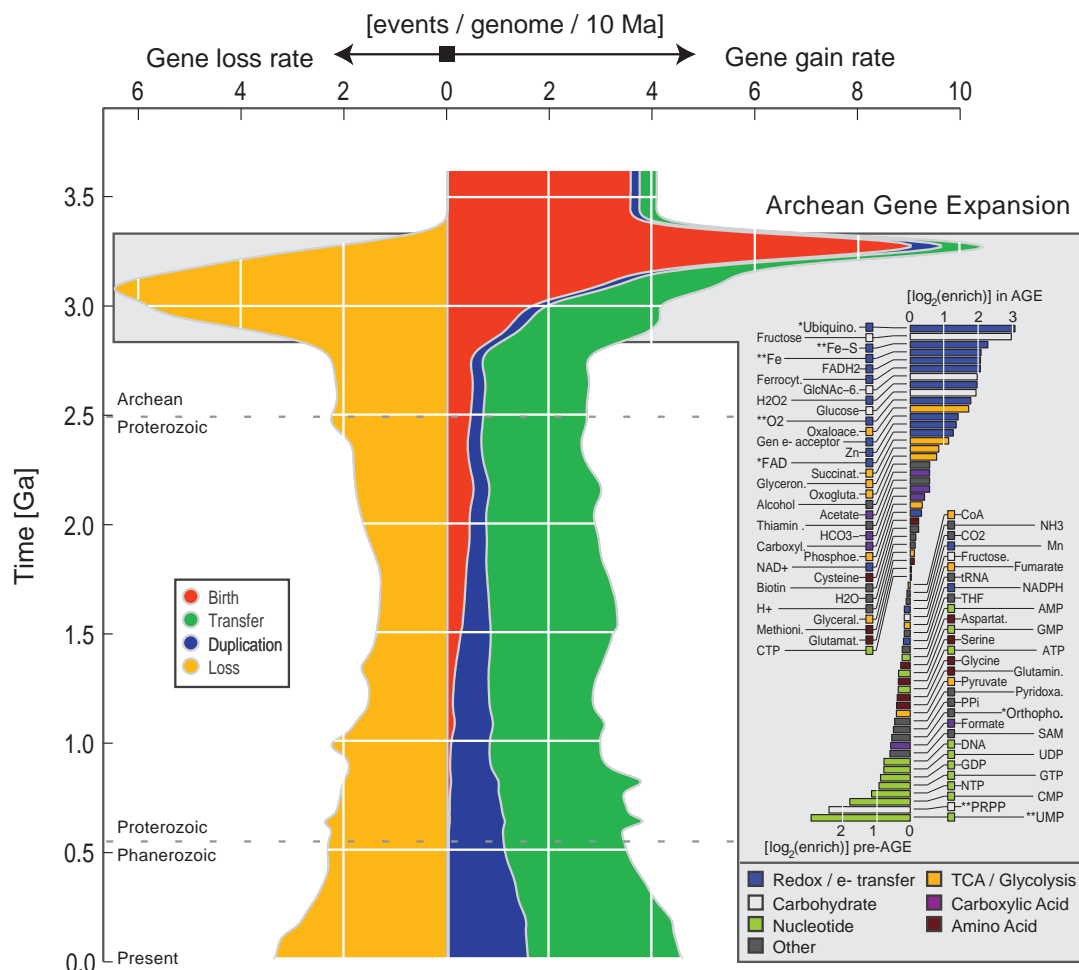


Figure 2.2: Rates of macroevolutionary events over time. The figure shows the rates of macroevolutionary events (colors as in Figure 2.1) as a function of time. Shown are the average rates of evolutionary events per lineage (events / 10 Ma / lineage). Events that increase gene count are plotted to the right, and gene loss events are shown to the left (yellow curve). Genes already present at the LUCA are not included in the analysis of birth rates because the time over which those genes formed is not known. The AGE was also detected when alternative chronograms were considered (Figure 2.12). Inset: metabolites or classes of metabolites ordered according to the number of gene families that use them that were born during the AGE compared to the number born before the expansion. Metabolites whose enrichments are statistically significant at a False Discovery Rate < 10% or < 5% (Fisher's Exact Test) are identified using one or two asterisks, respectively. Bars are colored by functional annotation or compound type (functional annotations were assigned manually). Metabolites were obtained from the KEGG database release 51.0 [117] and associated with COGs using the MicrobesOnline September 2008 database [118]. Metabolites associated with fewer than 20 COGs or sharing more than 2/3 of gene families with other included metabolites are omitted.

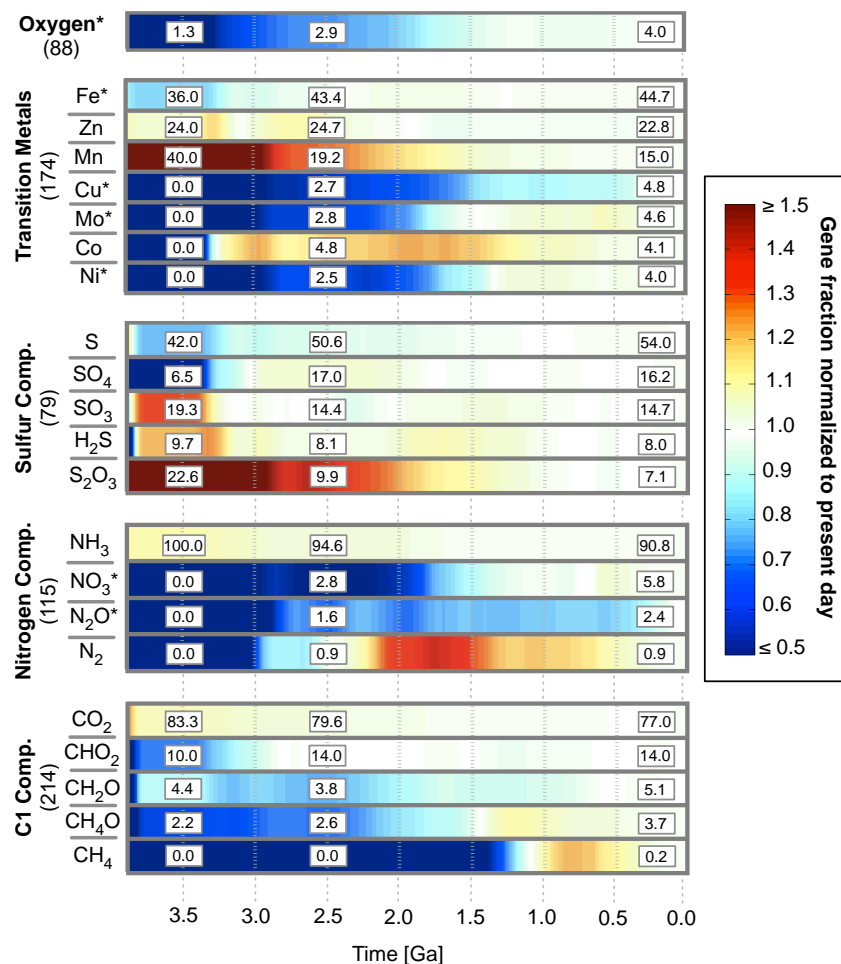


Figure 2.3: Genome utilization of redox-sensitive compounds over time. The first bar illustrates a gradual increase in the fraction of enzymes that bind molecular oxygen predicted to be present over Earth history ($p=1.3 \times 10^{-7}$, two-sided Kolmogorov-Smirnov test). Colors indicate abundance normalized to present-day values. The lower four panels group transition metals, nitrogen compounds, sulfur compounds, and C₁ compounds. The fraction of each group's associated genes that bind a given compound, normalized to present-day fractions, is shown over time using a color gradient. Enclosed boxes show raw fractional values at three time points: 3.5 Ga (left); 2.5 Ga (middle); and the present day (right). For example, 18.9% of transition metal-binding genes are predicted to have bound Mn at 2.5 Ga, a value 1.26 times the size of the modern day percentage of 15.0%. Values within parentheses give the overall number of gene families in each group. To determine which compounds showed divergent genome utilization over time, the timing of copy number changes for each compound's associated genes was compared to a background model derived from all other compounds. Compounds whose utilization significantly differs from the background model are marked with an asterisk (False Discovery Rate < 5%, two-sided Kolmogorov-Smirnov test). Nitrite and nitric oxide are not shown due to their COG-binding similarity to nitrate and nitrous oxide, respectively.

2.2 Supplementary Material

2.2.1 Overview

We developed a phylogenomic method that we named AnGST (Analyzer of Gene and Species Trees), which “reconciles” any observed differences between a gene tree and a reference tree (species tree) by inferring a minimal set of evolutionary events, including horizontal gene transfer (HGT), gene duplication (DUP), gene loss (LOS), speciation (SPC) and exactly one gene birth or genesis event (GEN). Each event type is assigned a unique cost, and the overall sum of costs associated with a reconciliation is minimized (i.e., we use a generalized parsimony criterion). We address previously described shortcomings of similar parsimony-based models of host-parasite evolution [119] by accounting for phylogenetic uncertainty (using a new approach described below) and directly estimating event costs from our large dataset. We divide the gene-tree/species-tree reconciliation process into two components:

- The **basic reconciliation** step assumes a known gene tree and species tree and identifies the set of evolutionary events (HGT, DUP, LOS, SPC, GEN) needed to explain any discordance between the trees
- The **tree amalgamation** step accounts for gene tree uncertainty by incorporating tree construction into the reconciliation process: multiple gene tree bootstraps are provided to AnGST and the algorithm retains and combines bootstrap subtrees which yield the most conservative reconciliation consistent with the sequence data.

The estimation of **event costs** from the input data is based on reducing large fluctuations in ancient genome sizes. This method is presented in Section 2.2.5 together with a **sensitivity analysis** for the resulting parameters.

The AnGST software package is implemented in the Python programming language and can be downloaded from: <http://almlab.mit.edu/ALM/Software/>.

2.2.2 Basic Reconciliation Algorithm

First assumptions

The **basic reconciliation** step requires a rooted, strictly bifurcating gene tree G and species tree S . Each tree is composed of a set of nodes linked to one another by a set of connecting edges. We assume that each node g in G can be mapped to a node s in S , a mapping we abbreviate as $g:s$. This mapping describes which (extant or ancestral) genome hosted a given (extant or ancestral) gene copy. Maps are known with certainty for extant genes, but must be inferred for ancestral gene copies.

Algorithm explanation

Our goal in gene/species tree reconciliation is to recover the optimal set of evolutionary events that explain any topological discordance between the gene and species trees. A brute-force search through all possible evolutionary histories is intractable, as the number of possible histories grows exponentially with increasing tree size [120]. However, for a given gene and species tree pair, there are only $|S|$ possible mappings for the root node of the gene tree, g_r . If the optimal reconciliation is already known for each possible mapping $g_r:s_r$, where s_r is a node in S , a new outgroup for the gene tree can be added (making g_r a child of the new root node g_n), and optimal reconciliations for the larger gene tree can be quickly computed using the following method:

1. For each possible pair of mappings $(g_r:s_r, g_n:s_n)$ where s_r and s_n are nodes in S
 - (a) Choose the most parsimonious explanation for how a gene copy in s_n descended into s_r .
 - (b) Concatenate this history to the known optimal reconciliation for $g_r:s_r$, to produce the optimal reconciliation for the $(g_r:s_r, g_n:s_n)$ pair
2. Identify optimal reconciliations for each mapping $g_n:s_n$ by selecting the minimal overall reconciliation cost associated with $(g_r:s_r, g_n:s_n)$ as s_r is varied over the

nodes of S .

Using the above method, the reconciliation problem can be formulated in a dynamic programming framework, yielding computational complexity that is a polynomial-time function of gene tree size. The AnGST program implements this algorithm as a post-fix traversal of the gene tree. At each node, reconciliations from child subtrees are combined in mini-reconciliations, which explain how the gene copy at g coalesced from two child copies c_1 and c_2 (i.e., whether HGT, speciation, or duplication occurred), assuming the mappings $g:s$, $c_1:s_1$, and $c_2:s_2$. This is repeated for each s , s_1 , and $s_2 \in S$. Mini-reconciliations return optimal duplication-loss or HGT scenarios if s is the last common ancestor of s_1 and s_2 , or if s is identical to either s_1 or s_2 . All other combinations of s , s_1 , and s_2 , yield mini-reconciliations that we refer to as complex scenarios. We include these scenarios in the pseudocode below to aid understanding of basic reconciliation design, but we do not provide a method for their solution since complex scenarios can be safely ignored without loss of reconciliation optimality (see Running Time discussion below). If g is a leaf node, mini-reconciliations are unnecessary since the true mapping from g to the species tree is known. Once all combinations have been evaluated, we retain the optimal reconciliation associated with each possible mapping of g to the species tree. Pseudocode for the reconciliation algorithm is provided on the following page in Python style.

Pseudocode:

```
% Main %
• Reconcile(gene_tree.root)
% Methods %
• define Reconcile(node):
    • child_1, child_2 = ChildNodes(node) %strictly bifurcating tree
    • if child_1 AND child_2 are null: %is a leaf node
        • for node_map in AllNodes(species_tree):
            • if node_map is KnownHostGenome(node):
                • node.reconciliation_cost(node_map) = 0 %correct answer is known for leaves
            • else:
                • node.reconciliation_cost(node_map) = maxint
        • return
    • Reconcile(child_1) %post-fix traversal
    • Reconcile(child_2)
    • for node_map in AllNodes(species_tree): %try all possible hosts for ancestor
        • for child_1_map in AllNodes(species_tree): %try all possible hosts for children
            • for child_2_map in AllNodes(species_tree):
                • events = MiniReconcile(node_map, child_1_map, child_2_map)
                • prior_events_1 = child_1.reconciliation_cost(child_1_map)
                • prior_events_2 = child_2.reconciliation_cost(child_2_map)
                • overall_cost = Cost(events + prior_events_1 + prior_events_2)
                • cost_matrix(node_map, child_1_map, child_2_map) = overall_cost
    • for node_map in AllNodes(species_tree):
        • node.reconciliation_cost(node_map) = Min(cost_matrix(node_map, :, :))
    • return

• define MiniReconcile(node_map, child_1_map, child_2_map):
    • % compute DupLoss scenarios
    • if node_map is ancestral to child_1_map AND child_2_map:
        • if node_map is last_common_ancestor of child_1_map AND child_2_map:
            • %% See Page46 for DupLoss pseudocode
            • duploss_events = DupLoss(node_map, child_1_map, child_2_map)
        • else:
            • duploss_events = ComplexScenario()
            • %ComplexScenario() not implemented -- see Methods Section 1.1.2 Running Time discussion for explanation
    • else:
        • duploss_events = maxint % impossible to reconcile with only dup-loss
    • % compute HGT scenarios
    • if node_map is child_1_map:
        • hgt_events = {HGT from node_map to child_2_map}
    • elif node_map is child_2_map:
        • hgt_events = {HGT from node_map to child_1_map}
    • else:
        • hgt_events = ComplexScenario()
    • return MinCost(hgt_events, duploss_events)
```

An example reconciliation:

An AnGST reconciliation of two simple, but discordant, gene and species trees is provided in Figure 2.4. Here, we assume that we know the true mappings from the leaves in G to S : $g_1:s_A$, $g_2:s_C$, $g_3:s_B$. Because AnGST uses a post-fix traversal of G and the mapping of G 's leaves to S is trivial, we first investigate how g_4 is mapped to nodes in S . We initialize the algorithm by assigning infinite reconciliation cost to leaf mappings which deviate from the known leaf mappings (e.g. $g_1:s_B$); thus, there is only one valid mapping for g_1 and g_2 .

In Scenario α , g_4 is mapped to s_A ($g_4:s_A$) and we infer one HGT event using the mini-reconciliation algorithm (since g_4 is mapped to the same lineage as one of its child nodes). Similarly, if we consider $g_4:s_C$, we infer one HGT from s_C to s_A (Scenario β). In the case of $g_4:s_E$ (Scenario γ), g_4 is mapped to the LCA of s_A and s_E and a duplication-loss scenario is invoked by the mini-reconciler. Other more complex scenarios exist (e.g., $s_4:s_D$), but these can be ignored without affecting overall reconciliation optimality (see Running Time section below). Once optimal reconciliations have been found for each possible $g_4:s$ mapping, AnGST recurses to g_5 and repeats the process. In the next mini-reconciliation, there are multiple valid $g_4:s$ mappings. Thus AnGST must iterate through prospective mappings for both g_5 and g_4 (although for the sake of illustrative simplicity, we only enumerate a fraction of these scenarios).

In the first mapping shown for g_5 ($g_5:s_D$, $g_4:s_A$, $g_3:s_B$), there is 1 SPC (since s_A and s_B are direct vertical descendants of s_D) and this cost is added to the 1 HGT already inferred in Scenario α , which resulted in $g_4:s_A$. For the combination ($g_5:s_C, g_4:s_C, g_3:s_B$), a cost of 1 HGT (because g_5 and g_4 share the same mapping) is added to the cost for Scenario β . The last mapping shown is ($g_5:s_E, g_4:s_E, g_3:s_B$). A mini-reconciliation that posits HGT will imply forward-in-time gene transfers – an evolutionary event we do not allow (see *Temporal constraints on HGT* below). Instead, a DUP in s_E and subsequent losses among s_A and s_C are needed to correctly explain the mapping of $s_5:s_E$, $s_4:s_E$, and $g_3:s_B$. The g_5 mapping that leads to the

optimal reconciliation is a function of the chosen evolutionary event costs. With a cost structure: $C_{SPC}=0$, $C_{HGT}=1$, $C_{LOS}=2$, $C_{DUP}=3$, the optimal mappings would be $g_5:s_D$, $g_4:s_A$, and the associated reconciliation would be a GEN event at s_D , followed by SPC at s_D , and an HGT from s_A to s_C . However, if $C_{SPC}=0$, $C_{HGT}=10$, $C_{LOS}=2$, $C_{DUP}=3$, the optimal mapping would be $g_5:s_E$, $g_4:s_E$, and the associated reconciliation would be an initial GEN event at s_E , followed by a DUP in s_E , 2 SPCs each at s_E and s_D , and LOS in lineages s_A , s_B , and s_C .

Running time

$O(|G|^*|S|^3)$ is an upper bound on run-time complexity of AnGST, where $|S|$ and $|G|$ are the number of nodes in those trees, respectively. Running times can be significantly reduced without loss of reconciliation optimality, however, with a simple speedup. When performing mini-reconciliations on all combinations of $g:s$, $c_1:s_1$, and $c_2:s_2$ for s , s_1 , and $s_2 \in S$, any complex scenario (s is not s_1 or s_2 , and s is not the last common ancestor of both s_1 or s_2) will require at least two HGT (one to s_1 and another to s_2), or one HGT to the last common ancestor of s_1 and s_2 followed by a duplication-loss scenario originating at that ancestor. These more complex scenarios will therefore always be suboptimal with respect to non-complex scenarios and their evaluation can be skipped during the reconciliation process. The resulting reduction in mapping search space lowers AnGST run-time complexity to $O(|G|^*|S|^2)$. When temporal constraints on HGT are enforced (see below), this speedup cannot be fully exploited, as nodes ancestral to s_1 and s_2 are potentially optimal values for s in HGT scenarios.

In practice, on 3.0Ghz single-cores with access to 8GB of memory, an AnGST run reconciling 100 bootstrap trees from one gene family against a reference tree of 100 species would take roughly: 0.1 minutes for gene trees with 10 leaves, 4 minutes for gene trees with 50 leaves, 13 minutes for gene trees with 100 leaves, 27 minutes for gene trees with 150 leaves, 37 minutes for gene trees with 200 leaves.

Temporal constraints on HGT

If provided a chronogram as a reference tree, AnGST will restrict the set of possible inferred gene transfers to only those between contemporaneous lineages. This feature eliminates the possibility of inferring multiple HGT events which are chronologically impossible [31]. Any non-zero chronological overlap is sufficient to allow transfers. But, if a gene transfer is inferred from node s_1 to node s_2 , subsequent transfers of the gene copy in s_2 may only occur with lineages which exist during the range $T_1 \cap T_2$, where T_1 and T_2 are the times spanned by the parent edges of s_1 and s_2 , respectively. A feature enabling transfers forward in time (which may represent “phantom transfers” from unsampled taxa [121]) has been built into AnGST, but remains off by default and was not used in our analyses.

Gene tree rooting

Bootstrap trees are assumed to be unrooted. All possible rootings of these bootstrap trees are evaluated during the reconciliation process. The resulting gene tree is rooted on the branch that results in the overall lowest reconciliation score.

2.2.3 Bootstrap tree amalgamation

Errors or uncertainty in gene phylogenies can lead to the inference of spurious macroevolutionary events [32] and is a particular concern for deeply branching phylogenies [122]. AnGST resolves uncertainty by incorporating reconciliation into the tree-building process: the tree with the lowest reconciliation cost is chosen from a large ensemble of trees consistent with the sequence data. To generate an ensemble of suitable trees, AnGST considers the set of all trees that contains only bipartitions observed in a set of input trees, which we generate with non-parametric bootstrapping. Thus, AnGST typically outputs chimeric trees that do not match any of the input bootstrap trees exactly, although every bipartition in the AnGST tree occurs in at least one of the bootstraps. In simulations, we observe these trees to be significantly more accurate than trees based on sequence likelihood alone, although they

generally have lower likelihood (see *Chimeric tree fidelity* below and Figure 2.5). Any number of bootstrap trees can be used, but we found limited increase in accuracy in simulated data as a result of using more than 10 (data not shown).

We implement this approach in the following manner (see Figure 2.6 for an example). Given n gene tree bootstraps $\{G_1, G_2, \dots G_n\}$ and a reference tree S , AnGST will begin the basic reconciliation algorithm starting on tree G_1 . Each time AnGST evaluates an internal node g of G_1 , it also evaluates the set of internal nodes $I = \{g_1, g_2, \dots g_k\}$ in other bootstrap gene trees that define the same bipartition as g . The optimal reconciliation at this node is the lowest scoring scenario/topology observed in any of the bootstrap trees. That is, a distinct solution is computed for each possible mapping ($g_i:s$ for $g_i \in I, s \in S$), and only the best solution is retained for each value of s . These $|S|$ optimal mappings and their reconciliations are subsequently shared across all the nodes in I . This last step creates “chimeric” gene trees, as the reconciliation at g in G_1 may now refer to a topology found in bootstrap G_i .

2.2.4 Simulation and Benchmarking

Simulation

Benchmarking

We used simulations to benchmark the performance of AnGST. Ten independent gene trees births were simulated on each of the 199 extant and ancestral lineages of the reference tree. A simple Poisson statistics-based model of HGT, DUP, and LOS was used to generate random gene histories and associated gene trees; the average simulated gene family underwent 0.21 HGT, 0.05 DUP, 0.76 SPC, and 0.26 LOS per extant gene copy. (For comparison, our analysis of the COG dataset inferred 0.29 HGT, 0.10 DUP, 0.83 SPC, and 0.27 LOS per extant gene copy.) Synthetic amino acid sequences were generated using these simulated trees and the SeqGen software (v.1.3.2) [123]. Trees were reconstructed from the synthetic sequences using either the BIONJ algorithm (implemented in PhyML), PhyML (v.2.4.5) [72], or AnGST via 100 PhyML-generated bootstrap topologies (see Section 2.2.8 for PhyML parameters). A

subset of 75 gene families were used to learn costs for HGT and DUP (see Methods Section 2.2.5). A cost combination of $C_{HGT}=4, C_{DUP}=3$ minimized genome size flux using this gene family subset (compared to $C_{HGT}=3, C_{DUP}=2$ learned for the COG dataset).

Chimeric tree fidelity

Following reconciliation, nodes deep in the interior of the resultant gene tree can contain topologies not found in any of the inputted bootstraps (although all possible bipartitions of these subtrees will exist in at least one of the bootstraps). Thus, the potential search space of topologies is vast. We tested the fidelity of the chimeric gene trees learned during the reconciliation process using the Robinson-Foulds (RF) statistic [124], which measures the number of bipartitions not shared by a pair of trees. A 0 RF score indicates perfect concordance (all bipartitions of the candidate and reference tree are identical) and increasing RF scores denote higher phylogenetic discordance. Analysis of the 225 gene trees with a minimal level of complexity (more than 10 leaves) demonstrates that AnGST trees are significantly more accurate than trees generated by BIONJ ($p=9.110^{-8}$ Wilcoxon rank sum test) or PhyML alone ($p=1.810^{-2}$ Wilcoxon rank sum test). Interestingly, this increase in topological accuracy comes with a likelihood tradeoff in comparison to the PhyML algorithm ($p=2.710^{-39}$ Wilcoxon rank sum test). As an aside, we note that the PhyML likelihoods in these analyses are in agreement with previous simulations which showed PhyML capable of constructing trees with higher likelihood than the true topologies [72].

Inferred birth date accuracy

We benchmarked the accuracy of gene family birth dates predicted by AnGST using the 747 synthetic gene families that included more than one extant gene copy. A comparison of inferred birth events and the simulated age of birth events is shown in Figure 2.7A. There is a strong correlation between inferred and simulated ages (0.88) and 76% of births are predicted to within 250 My of their simulated age. These

results are especially promising given the noisy processes (sequence simulation and phylogenetic inference) separating simulation of a gene family and its reconciliation. Moreover, we see no obvious evidence of inference bias which may lead to the false inference of a birth spike. Direct comparison of birth counts during the AGE (2.9-3.3 Ga) to simulated births during the same period (Figure 2.7B) did not show a bias towards over-counting births. We did, however, observe a bias toward gene birth prior to the AGE, suggesting that our set of very ancient genes (born prior to 3.3 Ga) may be inflated.

2.2.5 Parameter learning

Minimizing genome size flux

We address the problem of assigning the costs to each event type in a manner similar to some previous studies [25, 27, 125]: we use predictions of ancestral genome sizes to constrain the costs C_{DUP} and C_{HGT} (these are the only free parameters as we can assume $C_{LOS}=1$ and $C_{SPC}=0$ without loss of generality). However, we chose to minimize differences in genome size between parent and child nodes (a metric we refer to as genome size flux) rather than constraining overall genome size over time for two reasons: first, gene acquisition rates may not have been constant over time and ancient genomes may have been smaller (or larger) than modern day genomes [125]; second, the extinction of ancient gene families would lead to a trend of smaller inferred ancestral genome sizes at earlier times even if actual genome sizes were constant. A grid search of cost space showed genome size flux to be minimized at: $C_{HGT}=3$ and $C_{DUP}=2$ (Figures 2.8A, 2.9).

Sensitivity analysis

We investigated the extent to which the high fraction of overall gene birth detected during the Archean Gene Expansion was dependent on model parameters (Figure 2.8B). Gene birth patterns were invariant over a broad range of C_{DUP} . Gene birth from 2.8-3.4 Ga dissipated only at low C_{HGT} values. However, this regime of C_{HGT}

resulted in unrealistic genome size distributions: ancestral genomes were much smaller than present day ones, and most genes were predicted to have been born on terminal branches and spread via HGT.

2.2.6 Reference tree construction

Building a Chronogram of Life

We used a previously reported Tree of Life as the template for a reference chronogram [90]. This template was constructed using a concatenation of 31 translation-related orthologs. All of the species represented in our gene family dataset were present in this template tree. Divergence times were estimated using PhyloBayes (v.2.3c) [126]. Since autocorrelated molecular clock models have been shown to outperform uncorrelated ones in some cases similar to this study [91], we ran PhyloBayes with a CIR process model of rate correlation. Eight sets of temporal constraints that could be directly linked to fossil or geochemical evidence were used and are displayed in Figure 2.10. Benchmarking PhyloBayes runs in parallel (n=95) established that predicted divergence times and model likelihood converged after a burn-in of roughly 1500 model cycles. Final divergence time estimates were estimated following a burn-in of 2500 cycles, after which trees were sampled every 20 cycles until the 3500th cycle.

2.2.7 Alternate reference trees

We tested the extent to which the AGE was sensitive to the topology of the reference tree and to the molecular clock model used in chronogram construction. We built 10 separate reference phylogenies using non-parametric bootstrapping of the Ciccarelli et al. gene alignment [90] (see Section 2.2.8 for PhyML parameters) and rooted each with either the Bacteria, the Archaea, or the Eukarya as the outgroup. Unequivocal errors in phylogeny that may be due to sequence alignment construction errors were observed for the *Bdellovibrio*, *Shigella*, *Treponema*, and *Helicobacter pylori* taxa; these were resolved by manual pruning and re-grafting. Each phylogeny was then converted to an

ultrametric tree using r8s (v.1.71) [116] under a penalized likelihood model (with an additive penalty function, truncated Newton nonlinear optimization, cross-validation enabled, a cross validation start value of 10, a cross validation smoothing increment of 3, and the number of smoothing values tried set to 4). The same set of temporal constraints was used as for PhyloBayes. For the purposes of computational economy, a subset of 250 COGs was randomly selected from our dataset and reconciled against each of the 10 alternate chronograms.

Predicted birth ages were robust to the usage of alternative chronograms. The median gene family birth date difference between any two alternative chronograms is 0.09 Ga (Fig. 2.11). Elevated rates of gene birth during the Late Archean were observed in all 30 of the alternative chronograms and on average, 19% of the 250 chosen COGs were predicted to be born during a 200-My window (Fig. 2.12). However, the timing of AGE-like window diverged from that reported by PhyloBayes and spanned 2.7-2.5 Ga (compared to 3.3-2.9 Ga for PhyloBayes). Inspection of the r8s chronograms suggests this temporal discrepancy may be related to differences in dating the cyanobacteria under the two models. For both the r8s and PhyloBayes chronograms, AGE-like events coincided with the relatively brief period during which the major bacterial phyla, such as the *Firmicutes* and *Proteobacteria*, diverged from the last bacterial common ancestor. This period of compressed cladogenesis predates the appearance of the cyanobacteria by roughly 100-200 My in both models. Our r8s analysis places the initial occurrence of the cyanobacteria at 2.5 Ga, which is precisely the minimum age constraint for the appearance of this clade (see Section 2.2.6). Using the same constraints, PhyloBayes predicts the cyanobacteria to have emerged 3.0 Ga. We confirmed the importance of the cyanobacteria in dating the AGE by reducing the minimum constrained age of this clade during r8s chronogram construction; the resultant model yielded a younger AGE (data not shown).

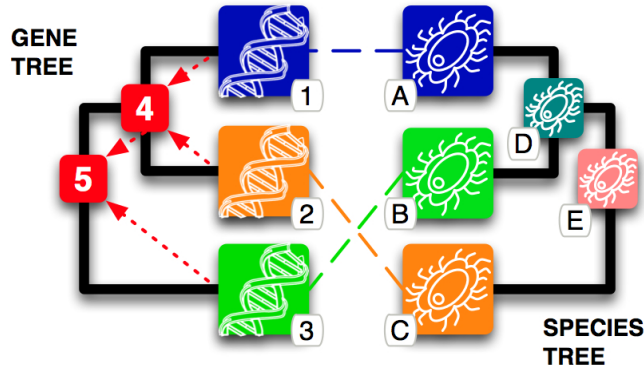
2.2.8 Gene tree construction

Families of orthologous genes used in this study are based upon functionally annotated orthologous groups from the COG database [127], as extended to a wider set

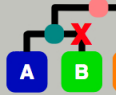
of genomes in the eggNOG database [106]. Due to computational limitations, we restricted this study to a subset of 100 of these genomes (11 eukaryotic, 12 archaeal, and 67 bacterial) broadly distributed across the Tree of Life. Sequences were downloaded from the eggNOG database in September of 2008.

eggNOG-derived families were filtered to ensure usable levels of sequence conservation with the aim of excluding the most error-prone phylogenies. We performed this filtering in an iterative fashion: First, we excised poorly aligned regions of sequence [90, 128], using Gblocks (0.91b) [129] with the minimum number of sequences for a flank position set to half the number of sequences in the alignment, the maximum number of contiguous non-conserved positions set to 8, the minimum length of a block set to 2, and the allowed gap positions set to all. Second, we excluded genes with more than 20% of their sequence in these excised regions from each gene family. Third, Muscle (v3.7) [130] was used with default settings to realign the remaining sequences. This process then returned to the first step, unless no sequences or regions were removed in the first or second steps in which case the process terminated. Of the original 4872 COGs, 788 lost more than 25% of their original gene copies during this process; these COGs were considered likely to be error-prone and thus excluded from further analysis. Another 101 COGs were not analyzed due to their high gene copy numbers and the extreme computational demands of running AnGST on those large families. A distribution of gene copy numbers within each gene family is shown in Figure 2.13.

Phylogenetic trees were constructed for the remaining gene families using version 2.4.5 of PhyML [72] and the following parameters: 100 bootstrap trees, a JTT substitution model, 0.0 percentage of the sites were invariable, 4 substitution rate categories, a gamma distribution parameter of 1.0, a BIONJ-based starting tree, and both tree topology and branch length optimization were enabled.



4

| CHILDREN | EVENTS | HOST | COST | |
|----------|--|------|-------|----------|
| A A | A → C | A | 1 x T | α |
| A C | C → A | C | 1 x T | β |
| A C |  | E | 1 x L | γ |

5


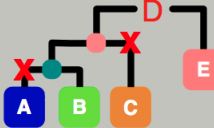
| CHILDREN | EVENTS | HOST | COST |
|----------|---|------|---------------------------|
| A B |  | D | 1 x S + α |
| C B | C → B | C | 1 x T + β |
| E B |  | E | 1 x D + γ 2 x L |

Figure 2.4: Example of a basic reconciliation. An AnGST reconciliation of two simple, but discordant, gene (G) and species (S) trees is shown. The mapping of leaves of G to S : $g_1:s_A$, $g_2:s_C$, $g_3:s_B$ is indicated with color (e.g., g_1 and s_A are both shown in blue). Reconciliation proceeds in a post-fix manner through the gene tree, first evaluating possible mappings from g_4 to nodes in the S . Once the reconciliation process is completed at g_4 , the algorithm continues at g_5 . A detailed explanation of this reconciliation is provided in Section 2.2.2.

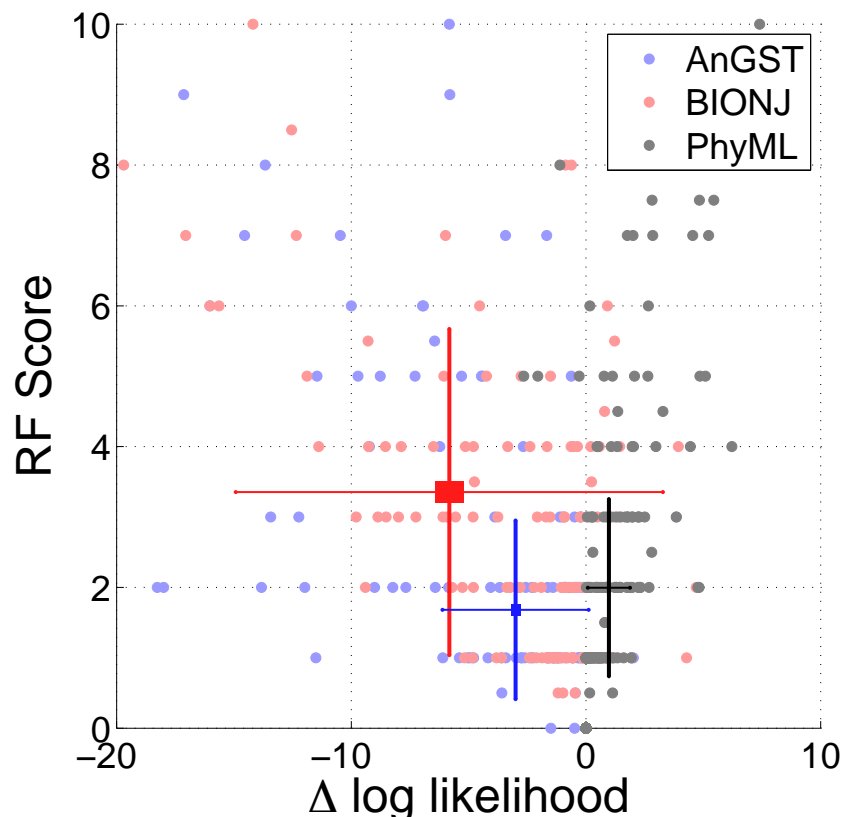


Figure 2.5: AnGST trees are more accurate than likelihood trees in simulation studies. We simulated the evolution of sequence data using 225 randomly generated gene trees with more than 10 leaves. Gene trees were reconstructed from synthetic sequence data using either BIONJ (red), PhyML (black), or AnGST (blue). Phylogenetic accuracy was evaluated by Robinson-Foulds (RF) score. A 0 RF score indicates perfect concordance (all bipartitions of the candidate and reference tree are identical) and increasing RF scores denote higher phylogenetic discordance. The logarithm of sequence likelihood given each tree model, relative to the likelihood calculated with the true gene topology, is plotted on the X axis. Mean RF scores and relative log likelihoods are drawn with rectangles whose height and width reflect standard errors of the mean; protruding lines are standard deviations. PhyML-based trees enjoy significantly higher likelihood scores than the AnGST chimeric trees ($p = 2.7 \times 10^{-39}$ Wilcoxon rank sum test), but the AnGST-based trees are significantly more similar to the correct gene tree topologies ($p = 1.8 \times 10^{-2}$ Wilcoxon rank sum test). Outlying points beyond axes were not drawn to facilitate viewing mean values, but were included in mean and significance estimations.

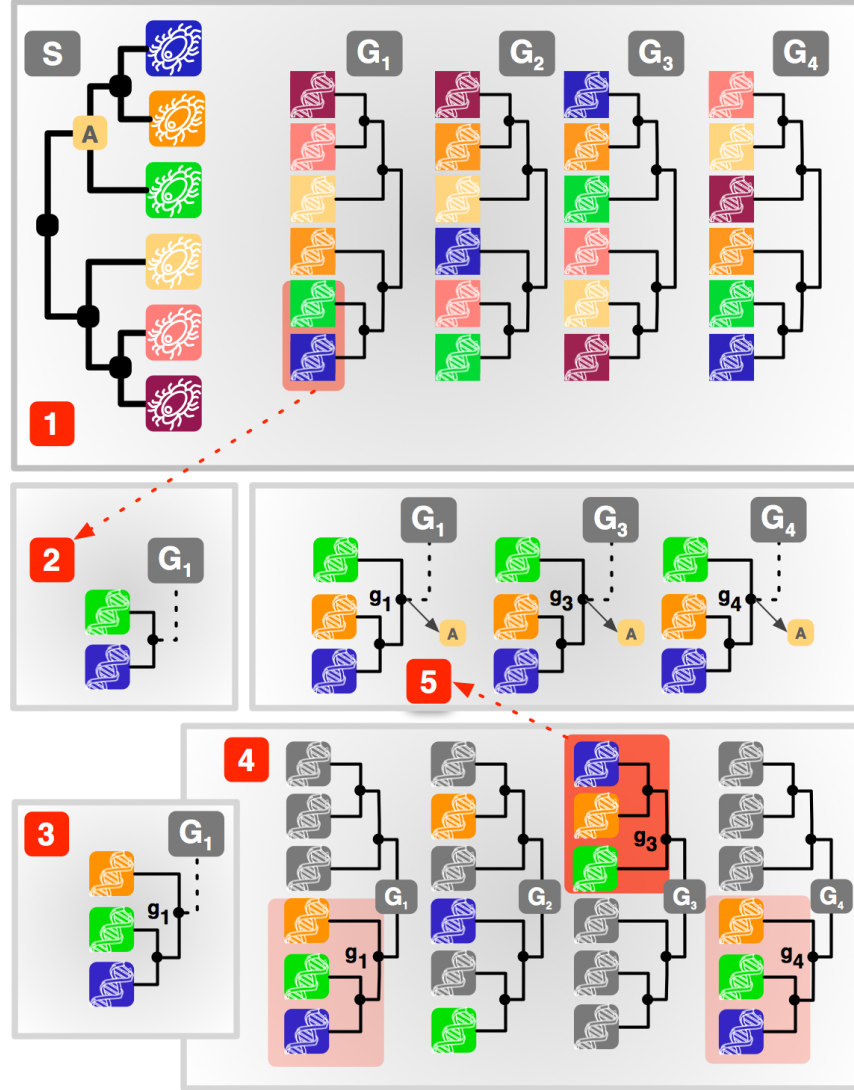


Figure 2.6: Amalgamation algorithm for phylogenetic uncertainty. An AnGST reconciliation of four gene tree bootstrap topologies $\{G_1, G_2, G_3, \text{ and } G_4\}$ and species tree S is shown. Leaf nodes on each bootstrap map to leaves on S according to color. The reconciliation begins on one of the bootstrap trees, G_1 (Step 1) and proceeds to an interior node (Step 2). The reconciliation does not consider other topologies for this subtree, as it only contains two leaves. When the reconciliation reaches the parent node node g_1 (Step 3), AnGST considers subtrees from other bootstraps with alternative topologies (but identical leaves). Corresponding subtrees are found on G_3 and G_4 and rooted at nodes g_3 and g_4 respectively (Step 4). Reconciliations are performed in parallel at g_1, g_3 , and g_4 . For the mapping of these internal nodes to lineage A on the species tree, the reconciliation at g_3 is optimal (since its topology matches the reference one) and the corresponding subtree in G_3 is substituted for the mappings $g_1:s_A$ and $g_4:s_A$.

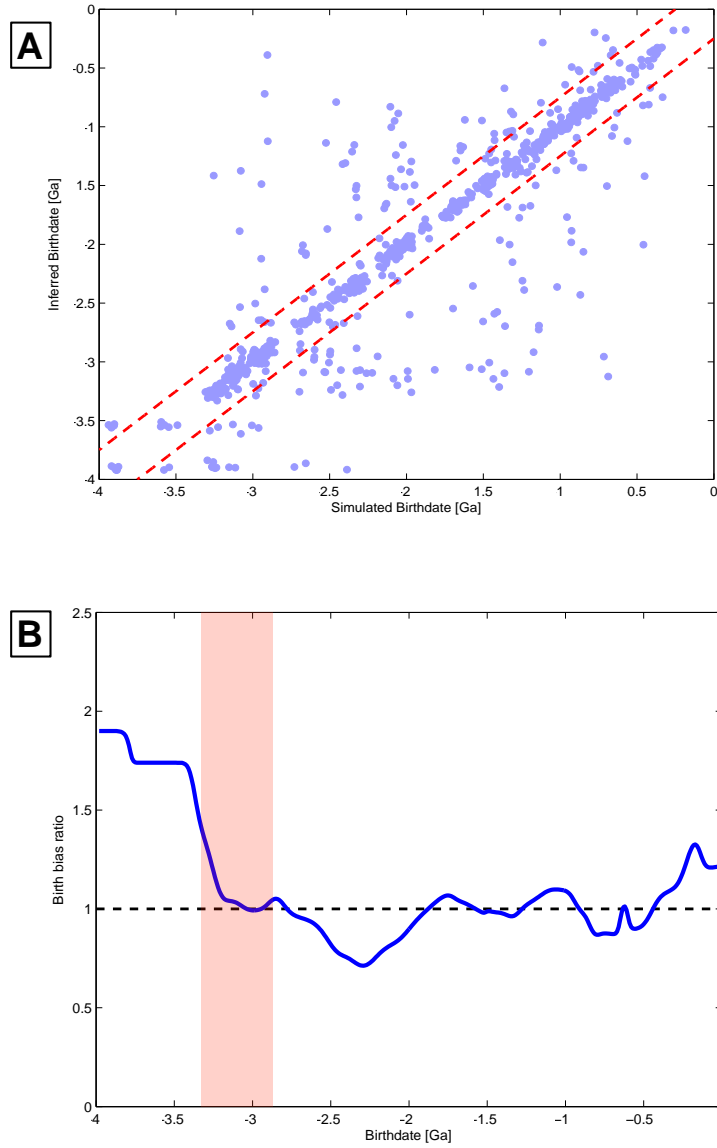


Figure 2.7: Benchmarking AnGST inference accuracy. A) A scatter plot of simulated gene family birth dates and inferred birth dates. Points drawn signify midpoints of branches associated with birth events. A slight amount of Gaussian noise with distribution $N(\mu=0, \sigma=0.025)$ has been added to each point so that overlapping points can be distinguished. The correlation coefficient is 0.88 and 76% of predicted births are within 250 My (bounded by red dashed lines) of their true ages. B) Birth prediction bias is plotted as a function of time. Predicted births have been normalized by the number of simulated births associated with a given age. The AGE (2.9-3.3 Ga) is highlighted in pink.

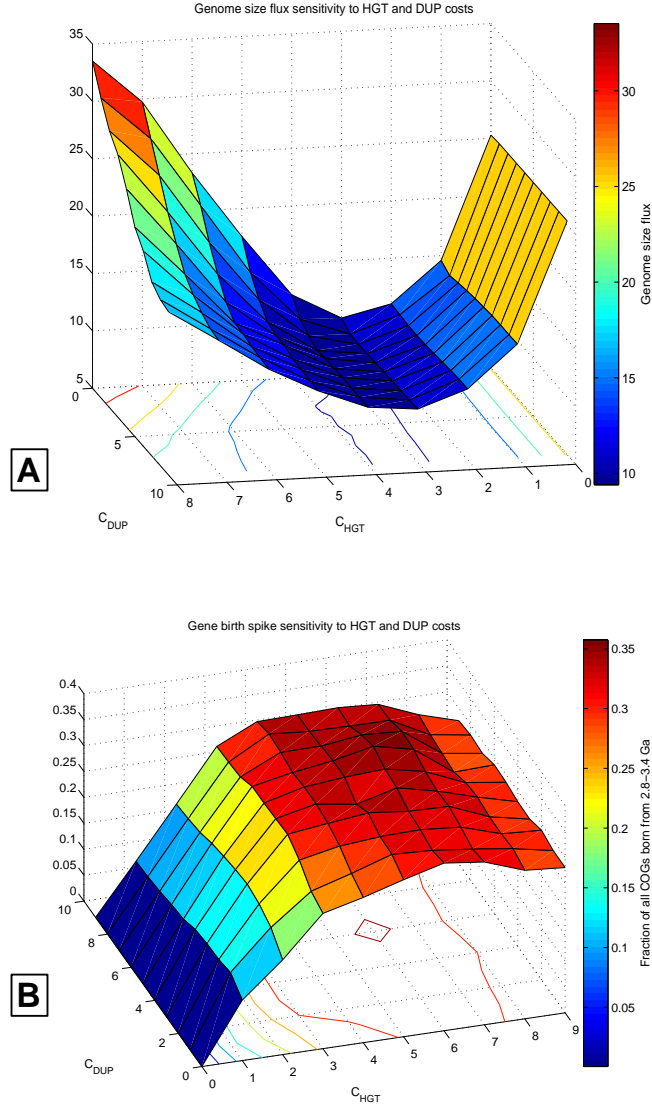


Figure 2.8: AnGST parameter learning and sensitivity analysis. A) We performed a grid search over the costs C_{HGT} and C_{DUP} with the intention of minimizing average genome size flux between inferred ancestral genomes. The costs C_{LOS} and C_{SPC} were fixed at 1.0 and 0.0, respectively. Flux can clearly be minimized along the C_{HGT} axis, but is less sensitive to changes in C_{DUP} . A minimum point does exist, however, at $C_{HGT}=3.0$, $C_{DUP}=2.0$. B) A sensitivity analysis for our detection of a high fraction of births from 2.8-3.4 Ga was performed over the same parameter space evaluated in A). Comparable fractions of overall gene birth to the AGE were detected in parameter space near the genome size flux minimum. Note that axes are reversed in panels A and B in order to facilitate viewing parameter sensitivity landscapes.

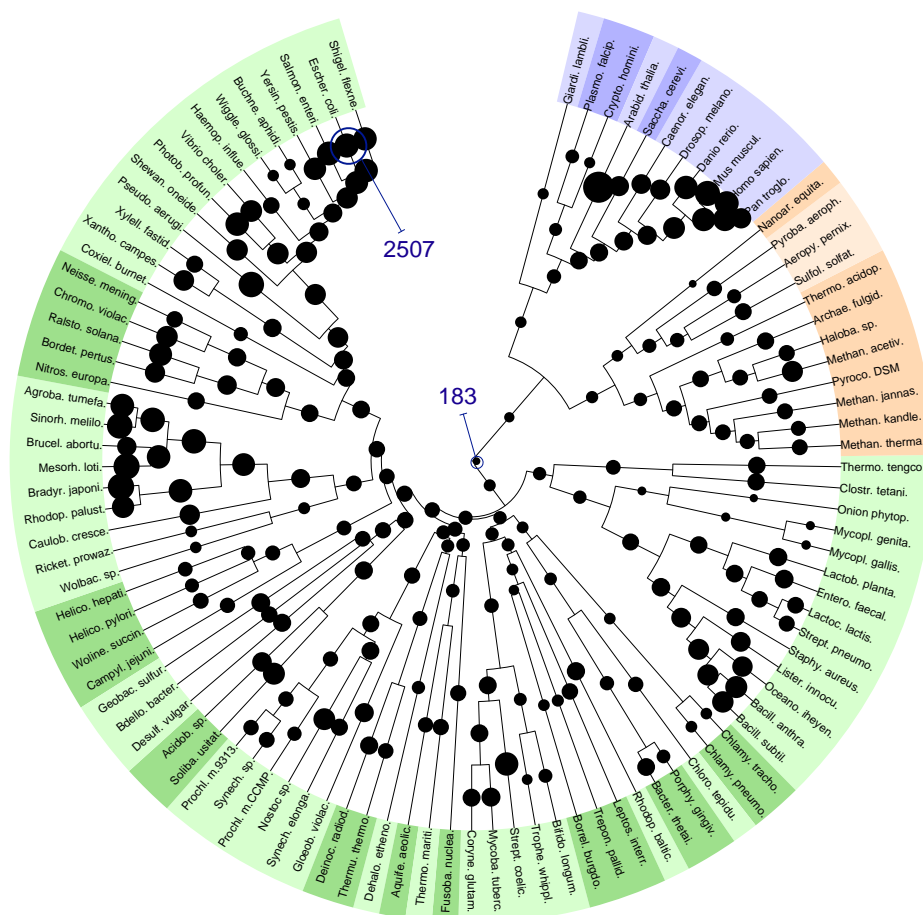


Figure 2.9: Inferred ancient genome sizes for $C_{HGT}=3.0$, $C_{DUP}=2.0$. Circle areas scale absolutely with genome sizes. Our optimization metric, genome size flux, aims to minimize the average difference between parent and child genomes. Ancestral genomes are predicted to be smaller than modern day ones; this may reflect the evolution of increasingly complex genomes, and/or the extinction of ancestral gene families. Genome sizes for the LUCA (183 genes) and the modern-day genome of *E. coli* (2507 genes) are labeled in dark blue. Metazoan genomes appear only slightly larger than prokaryotic ones because the COGs used in this study were originally defined using only unicellular organisms Tatusov 2000, which thus biased our analyses of eukaryotic genomes towards only microbially-related genes.

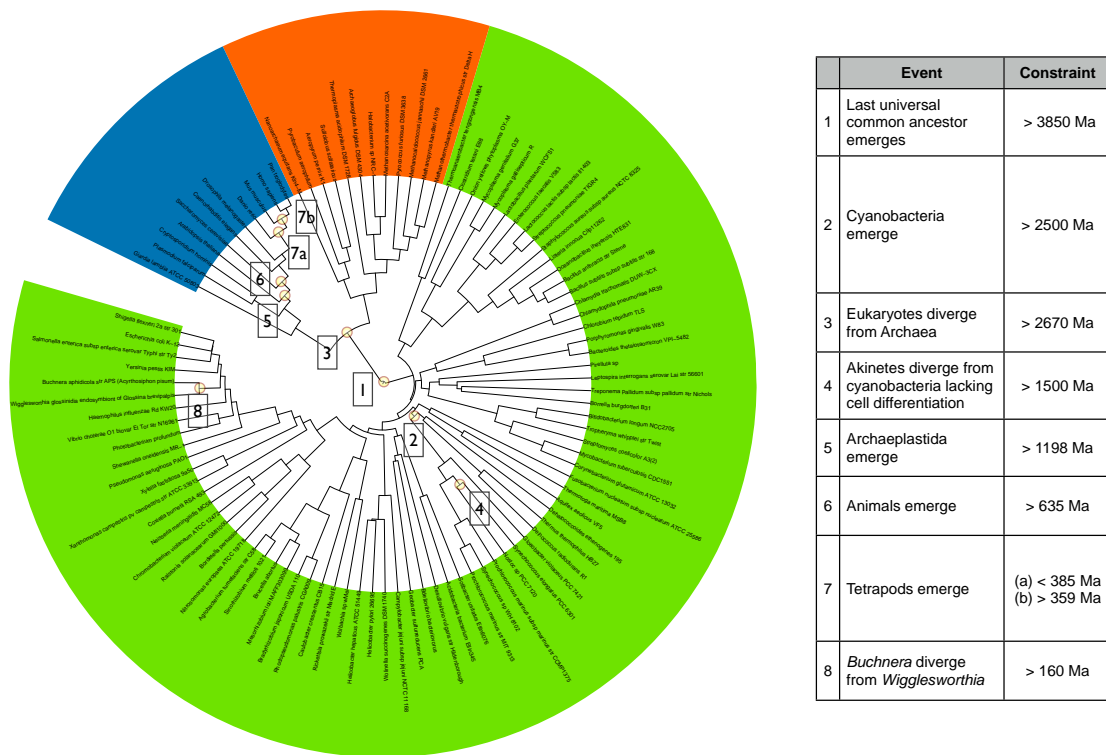


Figure 2.10: Temporal constraints. Eight fossil and biogeochemical constraints were used to constrain the chronogram (evidence cited can be found in Table 2.1). Those constraints are overlaid onto the reference phylogeny.

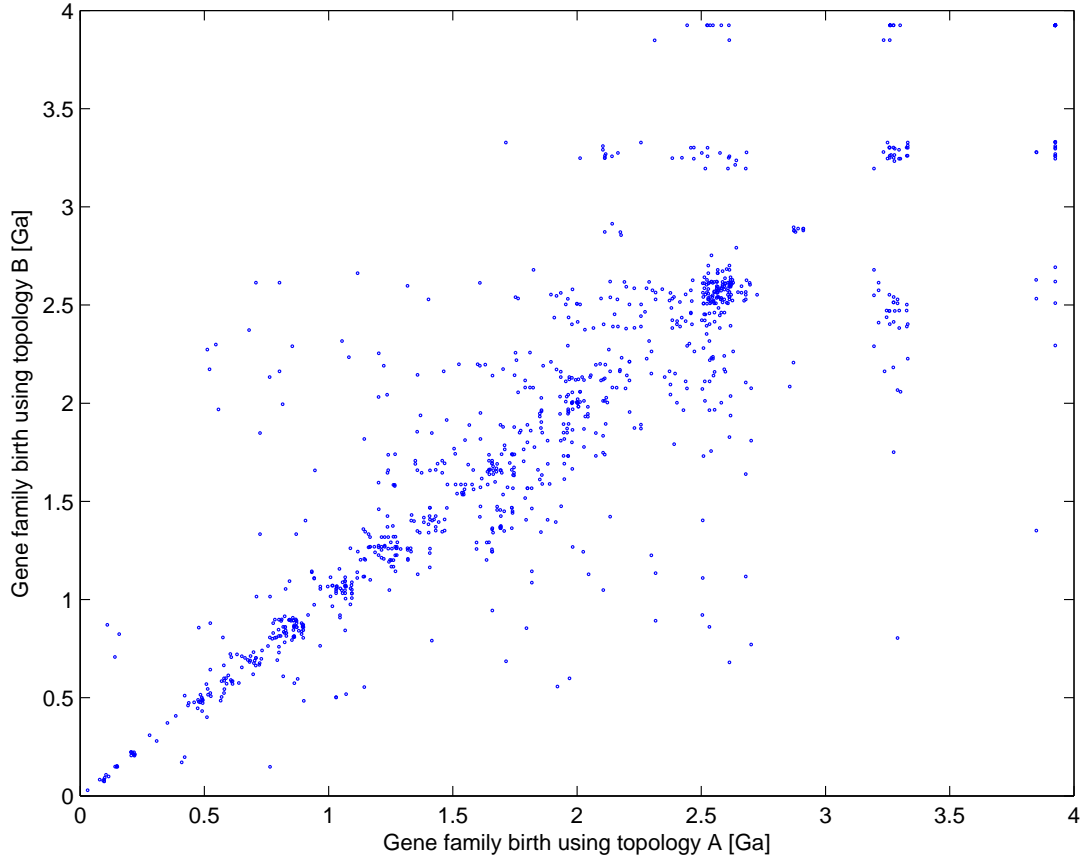


Figure 2.11: Sensitivity of predicted birth ages to variation in reference tree topology. Ten bootstraps of the reference tree were rooted using either the Bacteria, Archaea, or Eukarya as an outgroup and subsequently processed with r8s, producing 30 alternative reference chronograms. Birth dates were inferred for 250 gene families using each chronogram. We graph 1000 random combinations of alternative chronogram pairs and gene families in the scatter plot above (correlation coefficient = 0.86). The median gene family birth date difference between any two alternative chronograms is 0.09 Ga.

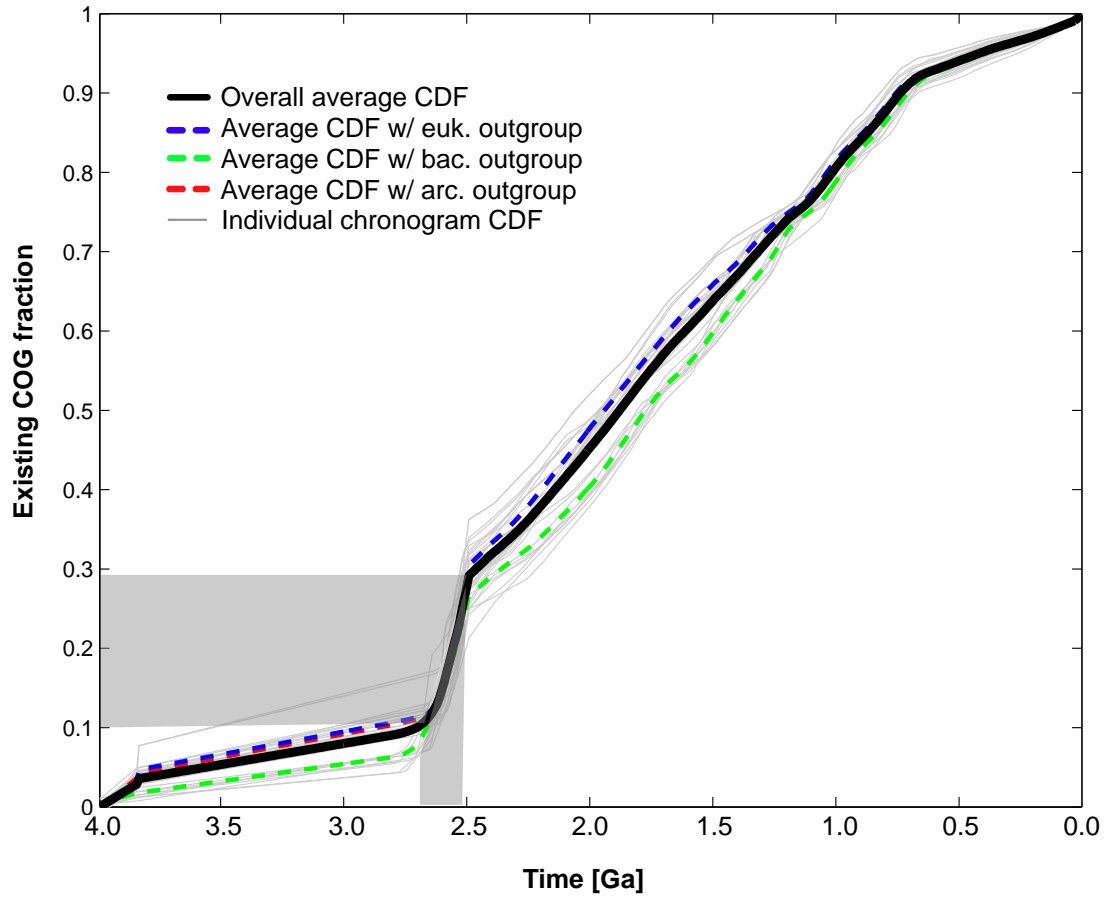


Figure 2.12: Gene family birth using 30 alternative reference tree topologies. Shown above are cumulative distribution functions (CDFs) of total COG birth over time for the 30 alternative reference chronograms (light gray lines). Mean CDFs for the Bacteria, Archaea, and Eukarya as outgroups are shown using green, red, and blue dashed lines, respectively. Overall (solid black line), the period 2.7-2.5 Ga witnesses a gene family birth spike of on average 0.23 families born per 1 Ma and accounts for the birth of 19% of the COG families studied. By contrast, birth rates average 0.07 families born per 1 Ma from 2.5 Ga-present day.

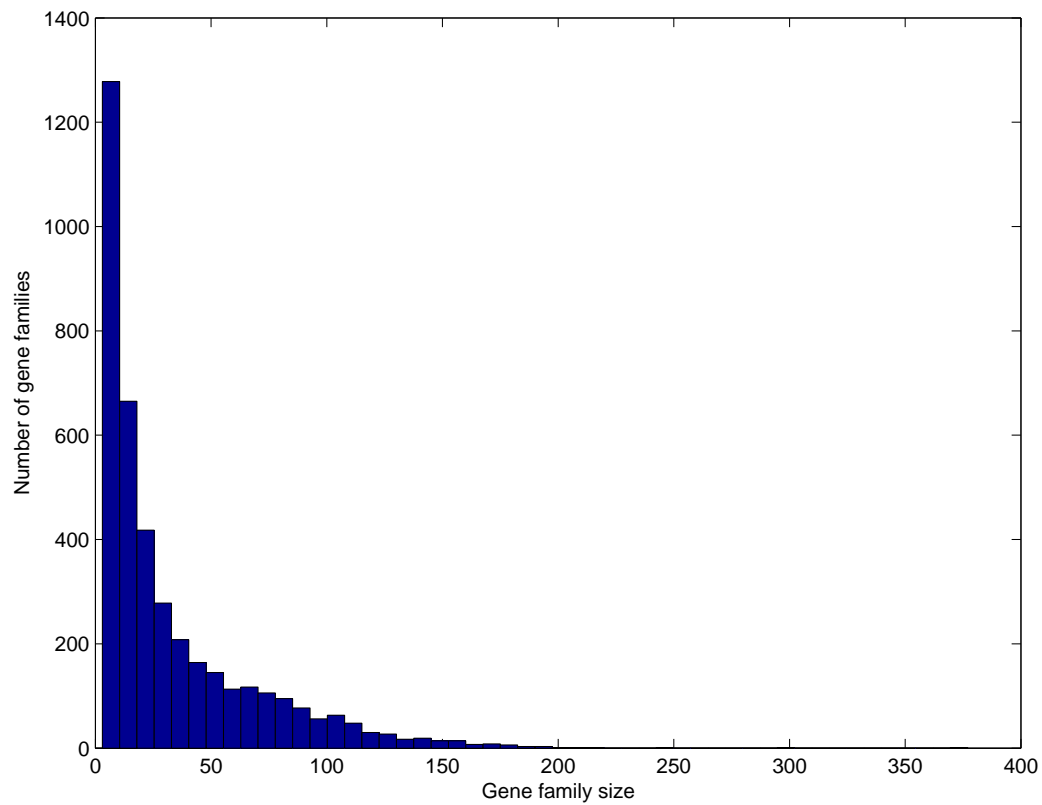


Figure 2.13: Histogram of COG family sizes. The median COG family in our dataset possesses 18 gene copies, and 93% of COG families have 100 or fewer gene copies.

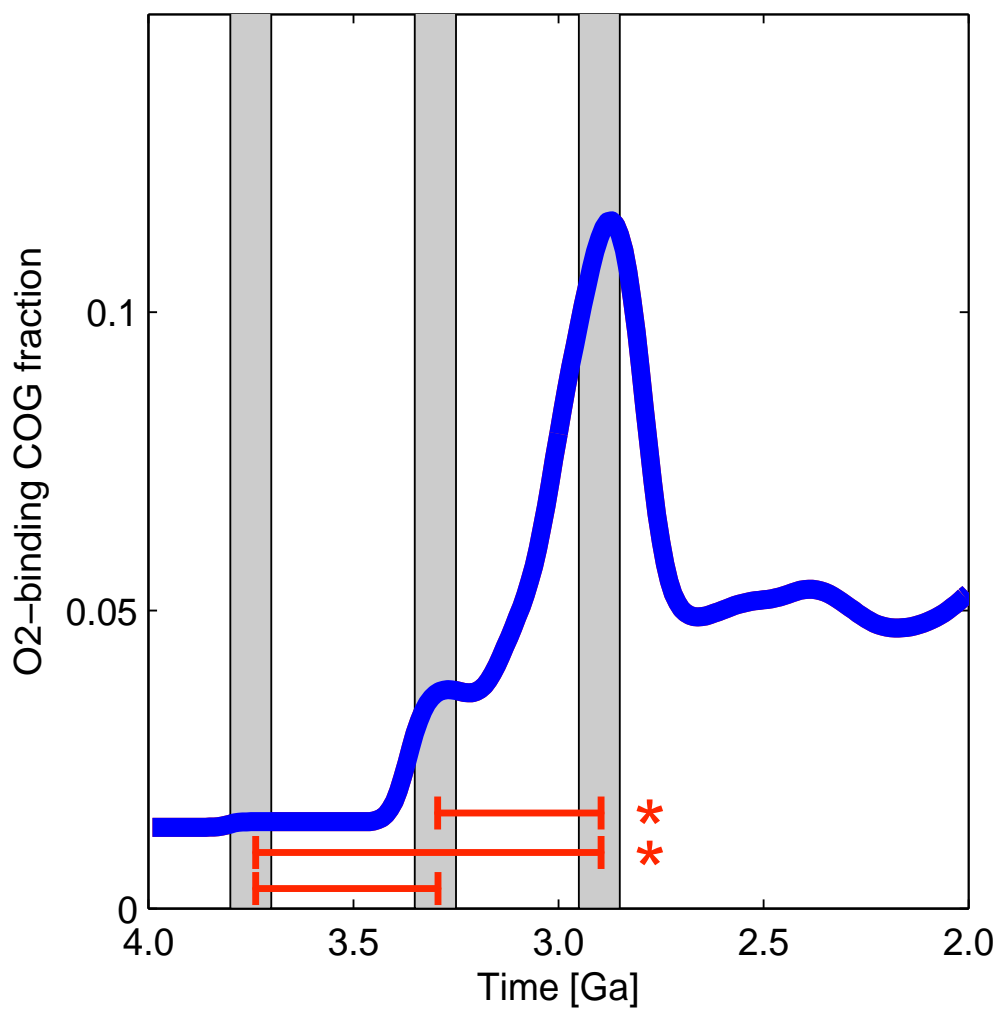


Figure 2.14: O₂-utilizing gene birth over time. The fraction of compound-binding COG births which bind O₂ is shown over time. A chi-square test was used to compare the overall number of COGs born and the number of O₂-binding COGs born in 100 My windows: prior to the AGE (3.7 Ga), at the height of the AGE (3.25 Ga), and at the tail of the AGE (2.85 Ga). Comparisons with $p < 0.05$ are denoted with asterisks on the graph. These data suggest that changes in O₂ usage came toward the end of the AGE.

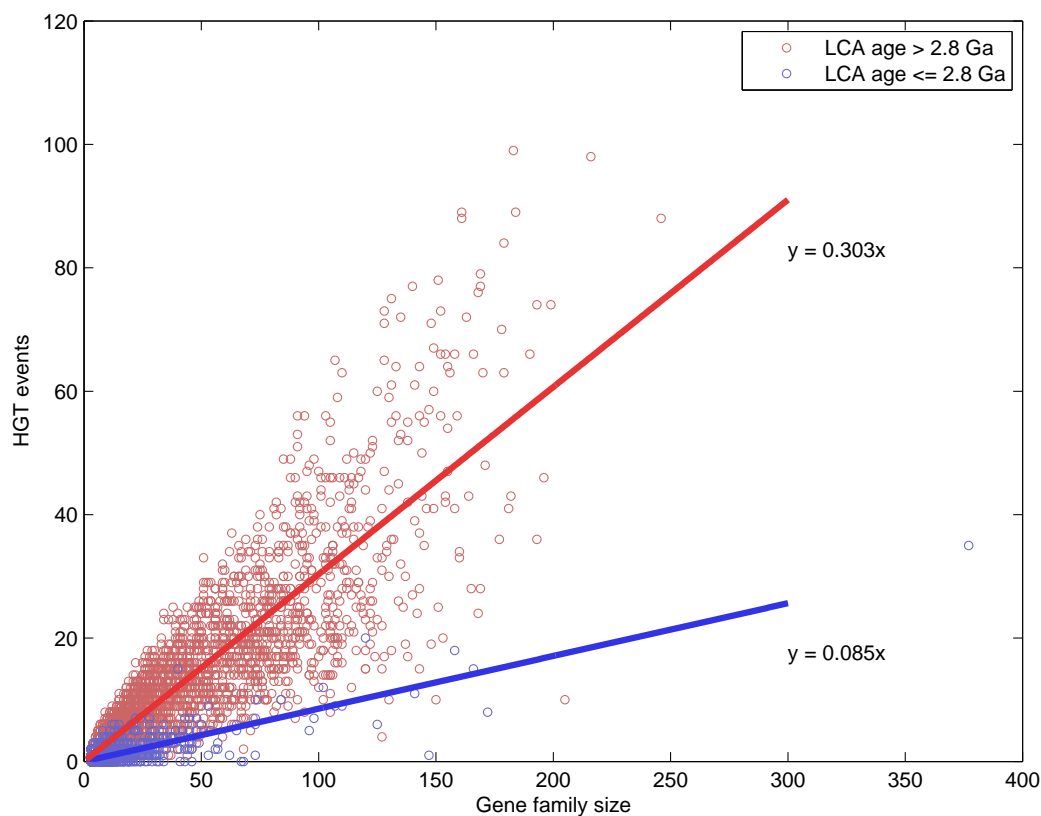


Figure 2.15: HGT counts vs. gene family size. The average gene family reconciliation yields 9.7 inferred HGT events. The number of HGT events inferred grows with the number of gene copies in a COG family. Gene family HGT counts also grow with the age of the last common ancestor of all genomes represented in the family, suggesting that HGT is more frequent among gene families spanning wider phyletic range. We note that y-intercepts for the above line fittings have been forced to equal 0.

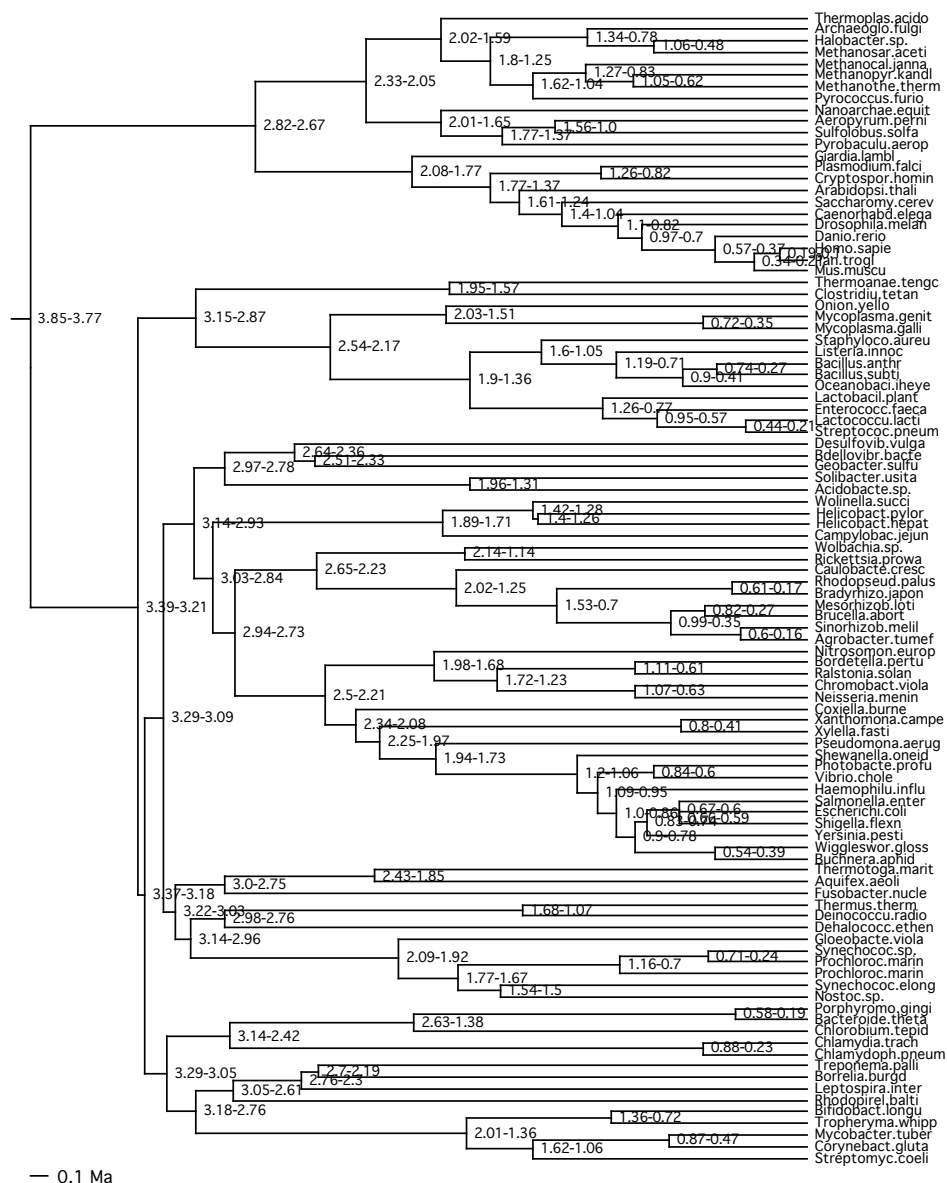


Figure 2.16: Confidence intervals for divergence times on reference chronogram. Confidence intervals (95%) were estimated by PhyloBayes and are shown next to each divergence point on the tree. Values are in units of Ga.

| Meta-function | Function | COG Code | Number of COGs | Fraction of genes studied | Fraction of AGE births | AGE birth enrichment | Fraction of pre-AGE births | pre-AGE birth enrichment | pre-AGE vs. AGE p-value |
|----------------------------------|-----------------------------|----------|----------------|---------------------------|------------------------|----------------------|----------------------------|--------------------------|-------------------------|
| Information storage & processing | Translation | J | 197 | 0.049 | 0.061 | 1.234 | 0.150 | 3.042 | 0.000 |
| | RNA proc. | A | 17 | 0.004 | 0.001 | 0.219 | 0.002 | 0.377 | 1.000 |
| | Transcription | K | 173 | 0.043 | 0.030 | 0.681 | 0.042 | 0.962 | 0.246 |
| | Replication, recombination | L | 155 | 0.039 | 0.034 | 0.868 | 0.068 | 1.746 | 0.002 |
| | Chromatin struct. | B | 11 | 0.003 | 0.000 | 0.000 | 0.002 | 0.655 | 0.338 |
| Cellular processes & signaling | Cell cycle control | D | 56 | 0.014 | 0.013 | 0.903 | 0.012 | 0.854 | 1.000 |
| | Defense mech. | V | 29 | 0.007 | 0.008 | 1.083 | 0.001 | 0.097 | 0.033 |
| | Signal transduction | T | 106 | 0.027 | 0.028 | 1.044 | 0.020 | 0.762 | 0.405 |
| | Cell wall/membrane | M | 141 | 0.035 | 0.050 | 1.424 | 0.060 | 1.702 | 0.487 |
| | Cell motility | N | 82 | 0.021 | 0.031 | 1.493 | 0.015 | 0.718 | 0.064 |
| | Cytoskeleton | Z | 5 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| | Intracell. trafficking | U | 122 | 0.031 | 0.032 | 1.047 | 0.022 | 0.710 | 0.274 |
| | Post-trans. modification | O | 157 | 0.039 | 0.043 | 1.101 | 0.042 | 1.070 | 1.000 |
| Metabolism | Energy prod. & conv. | C | 211 | 0.053 | 0.079 | 1.499 | 0.068 | 1.289 | 0.490 |
| | Carb. trans. & met. | G | 186 | 0.047 | 0.056 | 1.205 | 0.051 | 1.087 | 0.730 |
| | Amino acid trans. & met. | E | 226 | 0.057 | 0.079 | 1.394 | 0.109 | 1.923 | 0.054 |
| | Nucleotide trans. & met. | F | 83 | 0.021 | 0.026 | 1.228 | 0.059 | 2.835 | 0.001 |
| | Coenzyme trans. & met. | H | 155 | 0.039 | 0.056 | 1.439 | 0.069 | 1.772 | 0.326 |
| | Lipid trans. & met. | I | 72 | 0.018 | 0.024 | 1.306 | 0.032 | 1.795 | 0.334 |
| | Inorganic ion trans. & met. | P | 182 | 0.046 | 0.058 | 1.276 | 0.046 | 0.998 | 0.298 |
| | Secondary metabolites | Q | 70 | 0.018 | 0.015 | 0.847 | 0.004 | 0.216 | 0.045 |
| Poorly characterized | Func. unknown | S | 1186 | 0.298 | 0.183 | 0.615 | 0.071 | 0.238 | 0.000 |
| | General func. pred. | R | 560 | 0.141 | 0.142 | 1.010 | 0.113 | 0.804 | 0.105 |

Figure 2.17: Function of gene births prior to and during the AGE. Functional enrichment of gene birth from 2.8-3.3 Ga is shown for the 20 COG functional categories. A two-tailed Fisher exact test was used to compute the p-value of a difference in total COG births prior to the AGE vs. during the AGE, for each functional category (last column).

| Meta-function | Function | COG Code | HGT out of cyano into plant | HGT out of cyano | Fisher p-value | HGT out of alpha-proteos into euks | HGT out of alpha-proteos | Fisher p-value |
|----------------------------------|-----------------------------|----------|-----------------------------|------------------|-----------------|------------------------------------|--------------------------|-----------------|
| Information storage & processing | Translation | J | 38 | 1225 | <u>4.41E-05</u> | 13 | 779 | 9.34E-02 |
| | RNA proc. | A | 0 | 1 | 1.00E+00 | 0 | 0 | 1.00E+00 |
| | Transcription | K | 2 | 297 | 3.34E-01 | 0 | 231 | 1.78E-01 |
| | Replication, recombination | L | 4 | 550 | 1.52E-01 | 0 | 324 | 8.23E-02 |
| | Chromatin struct. | B | 0 | 7 | 1.00E+00 | 0 | 8 | 1.00E+00 |
| Cellular processes & signaling | Cell cycle control | D | 3 | 164 | 7.42E-01 | 0 | 98 | 6.28E-01 |
| | Defense mech. | V | 0 | 116 | 4.25E-01 | 0 | 59 | 1.00E+00 |
| | Signal transduction | T | 2 | 341 | 2.54E-01 | 3 | 209 | 4.84E-01 |
| | Cell wall/membrane | M | 5 | 735 | 6.06E-02 | 0 | 513 | 6.24E-03 |
| | Cell motility | N | 0 | 80 | 6.36E-01 | 0 | 168 | 4.23E-01 |
| | Cytoskeleton | Z | 0 | 0 | 1.00E+00 | 0 | 0 | 1.00E+00 |
| | Intracell. trafficking | U | 6 | 273 | 3.20E-01 | 1 | 305 | 3.79E-01 |
| | Post-trans. modification | O | 11 | 549 | 3.72E-01 | 9 | 367 | 1.55E-02 |
| Metabolism | Energy prod. & conv. | C | 26 | 949 | <u>3.95E-03</u> | 21 | 621 | <u>1.55E-06</u> |
| | Carb. trans. & met. | G | 9 | 541 | 7.21E-01 | 2 | 335 | 5.86E-01 |
| | Amino acid trans. & met. | E | 16 | 1205 | 6.23E-01 | 5 | 701 | 5.57E-01 |
| | Nucleotide trans. & met. | F | 6 | 486 | 8.49E-01 | 1 | 270 | 5.32E-01 |
| | Coenzyme trans. & met. | H | 13 | 1057 | 5.12E-01 | 7 | 397 | 1.98E-01 |
| | Lipid trans. & met. | I | 9 | 295 | 5.06E-02 | 4 | 247 | 3.33E-01 |
| | Inorganic ion trans. & met. | P | 12 | 918 | 6.76E-01 | 5 | 472 | 1.00E+00 |
| | Secondary metabolites | Q | 3 | 163 | 7.41E-01 | 0 | 106 | 6.30E-01 |
| Poorly characterized | Func. unknown | S | 16 | 1601 | 8.04E-02 | 5 | 1103 | 3.73E-02 |
| | General func. pred. | R | 19 | 1496 | 4.34E-01 | 9 | 797 | 7.17E-01 |

Figure 2.18: Biases in gene function associated with ancient endosymbioses. Shown here is a functional breakdown of HGT from cyanobacteria into *Arabidopsis thaliana*, and from the alpha-proteobacteria to ancient eukaryotes (which we defined as eukaryotic lineages predating the divergence of *Arabidopsis*). The significance of functional enrichments for genes associated with the chloroplast are calculated by comparing the observed number of HGT from cyanobacteria into the plant lineage, to the total number of HGT originating in the cyanobacteria, so as to account for functional biases associated with HGT out of the cyanobacteria. Similar statistics are performed on mitochondria-related genes. P-values that fall below a 5% False Discovery Rate cutoff in are underlined.

Chapter 3

Building prokaryotic species trees from thousands of gene trees

Lawrence A. David, Albert Y. Wang, Eric J. Alm

Experiments in this chapter were performed in collaboration with Albert Wang in the Alm laboratory.

Chapter 3

Building prokaryotic species trees from thousands of gene trees

3.1 Abstract

Sequenced-based approaches for reconstructing prokaryotic species trees from more than one gene utilize either the concatenation of sequences (supermatrices) or the merger of separate gene trees (supertrees) [131]. Yet, supermatrices ignore differences in evolutionary rate and nucleotide compositional bias between concatenated genes, which can ultimately reduce the accuracy of inferred phylogenetic trees [132]. Supertree methods can account for these inter-gene heterogeneities, but a commonly-used supertree technique, matrix representation, violates the site-independence assumption underlying many phylogenetic construction algorithms. Here, I extend an alternative supertree method known as Gene Tree Parsimony (GTP), which chooses the species tree as the topology that requires the least number of duplication events to be inferred when compared to a set of gene trees [133]. My GTP model, named GAnG, can account for horizontal gene transfer events (HGT) and is suitable for use with prokaryotic gene trees. Preliminary GAnG analysis of 250 archaeal gene trees built from a subset of the sequenced archaeal genomes supports hypotheses that Nanoarchaeota diverged from the last ancestor of the Archaea prior to the Crenarchaeota/Euryarchaeota split, but also appears sensitive to HGT artifacts between the

Crenarchaeota and the Thermoplasma.

3.2 Introduction

Prokaryotic species phylogenies are frequently built by concatenating sequences from more than one gene [131], in order to sample a range of phylogenetically-informative characters [134], and to mitigate the role of topological artifacts caused by long branch attraction [122, 135] and nucleotide composition bias [136, 137]. These genes can be restricted to “core” sets that are topologically congruent [90, 138], under the assumption that horizontal gene transfer (HGT) artifacts would only be present if identical HGTs affected all of the genes in the core set. However, reconstructions of organismal histories using core genes have been criticized for ignoring the majority of phylogenetic signal present in genomes and have been referred to as “The tree[s] of one percent” [139]. In order to claim that species trees represent overall genome evolution, computational tools are required that can incorporate genetic information from hundreds, or even thousands, of genes.

Previous studies that have built prokaryotic species trees from multiple gene families have utilized either “supertree” or “supermatrix” approaches [90, 138, 140, 141]. Under the supermatrix approach, sequences from orthologous gene families are concatenated into a single sequence, which can subsequently be inputted into a standard phylogenetic construction algorithm. This method is robust to gene families with limited taxonomic distribution, as simulations have shown supermatrices to perform well even in regimes where only 10% of species possess a given gene copy [142]. However, the concatenation of prokaryotic genes risks assembly of sequences with divergent evolutionary histories due to HGT [143]. Supermatrix approaches must therefore assume that the phylogenetic signal from vertically-descended genes overwhelms any signal from transferred regions of the alignment. Moreover, species trees cannot be constructed in a computationally-efficient way if differences in gene evolutionary rate or nucleotide composition are taken into account. Simulations have shown that partitioning concatenations and fitting these parameters on a per-gene basis improved the fidelity of reconstructed phylogenies in nearly all reported scenarios, and in some cases enabled the reliable recovery of clades that were never inferred by a conven-

tional supermatrix procedure [132]. However, supermatrices that model evolutionary heterogeneity between genes cannot utilize existing tree building algorithms and have so far required exhaustively searching tree space in order to identify an optimal tree [132, 134].

Supertree methods provide an alternative phylogenetic framework that can model evolutionary parameters on a per-gene basis, while still constructing a species phylogeny in a computationally-efficient manner. Under this framework, gene trees are built separately for each orthologous gene family and subsequently merged to form a species tree. Supertree methods are distinguished by how gene tree merger takes place. According to the most popular merger method [144], matrix representation using parsimony (MRP), a binary character matrix is built that possesses a column for each internal node in the set of gene trees, and a row for each of the species sampled. Species with a gene copy descended from a given internal node all share a “1” in the corresponding column of the matrix, and all other species share a “0” [145]. A consensus tree is formed by inputting the character matrix into a sequence-based phylogenetic construction algorithm. Because the supertree framework allows trees to be built for individual gene families, the model is capable of accounting for differential evolutionary rates and nucleotide compositions between genes, unlike conventional supermatrices.

However, the MRP algorithm in particular has been criticized for its analogy between the tree matrix and a nucleic acid sequence matrix [133]. This analogy violates the site-independence model assumption of many phylogenetic reconstruction algorithms [76], since sibling leaves on a gene tree will be partitioned similarly for each MRP matrix column that corresponds to one of these leaves’ ancestral nodes. In practice, this site dependence should manifest as a supertree bias towards subtree topologies from large gene trees with many ancestral nodes and therefore more sites in the tree matrix. This bias is likely deleterious, as large gene trees can be caused by frequent HGT and duplication events that impede partitioning gene families into orthologous groups.

Gene tree parsimony (GTP) provides an alternative supertree merger criterion

that has received less statistical criticism than MRP, but that also cannot be used on prokaryotic gene trees in present implementations. First introduced by Slowinski in 1997, the GTP approach chooses the species tree as the topology that requires the least number of duplication events to be inferred when species and gene trees are compared (a step called a reconciliation) [146]. This reconciliation-based approach avoids the matrix construction step that has led to criticism of MRP supertree methods. However, existing implementations of GTP can only model gene duplication and gene loss events [147, 148], limiting their usage to eukaryotic phylogenies.

Here, I enable the use of GTP supertrees on prokaryotic trees, through a new program that I have named GAnG. This algorithm reconciles gene trees and species trees using a generalized parsimony model I previously developed, the Analyzer of Gene & Species Trees, or AnGST [149], which accounts for HGTs, as well as gene duplication and gene loss events, in gene family evolution. Experiments with simulated data show that GAnG can accurately quantify species tree accuracy. As a proof of concept, I also used the algorithm to learn a phylogeny of 12 sequenced archaeal species from 250 gene trees with divergent topologies and that have likely undergone HGT. Results of this analysis support the basal position of *Nanoarchaeum equitans* on the archaeal tree.

3.3 Methods

3.3.1 Approach overview

The GAnG algorithm is an iterative process composed of four primary steps:

1. **Initial topology generation and reconciliation:** An initial reference tree is constructed using one gene, or a concatenation of multiple genes. Individual gene trees are also constructed for all orthologous gene families. After trees have been built, each of the gene trees is reconciled against the species tree using AnGST.
2. **Proposal of tree refinements:** The HGTs inferred from reconciliations between the species tree and gene trees are mined to propose subtree-prune-and-regraft (SPR) moves. These SPRs are used to generate candidate species tree topologies (Section 3.3.2). Alternatively, new rootings of the species tree can be evaluated (Section 3.3.3).
3. **Evaluation of candidate refinements:** Each candidate species tree is evaluated by comparison to the set of gene trees using AnGST, as described in Section 3.3.4.
4. **Selection of a new species tree:** The best scoring candidate tree is identified. If this tree has a better reconciliation score than the current species tree, the species tree is replaced with the candidate tree and the algorithm returns to Step 2. Otherwise, the process terminates and returns the present species tree.

3.3.2 Generating candidate species trees

GAnG searches species tree space using an iterative subtree-prune-and-regraft (SPR) strategy that generates a candidate species trees from an existing species tree by pruning off subtrees and regrafting them onto new locations on the tree. This heuristic approach is necessary because it is unlikely that a polynomial-time algorithm exists for finding an optimal species tree using AnGST [150]. An exhaustive search strategy

is also not possible, since the number of possible rooted binary trees grows super-exponentially with the number of species s , following the function $(2s - 3)!!$ [151] (for the 14 genomes in Fig. 3.2, there are nearly eight trillion possible organismal trees).

SPR moves enable dramatic changes in tree topology using relatively few topological operations and are used by tree construction algorithms like PhyML 3.0 [152], FastTree [153] and RaxML [154] to avoid being caught in local minima in tree space. An SPR-based strategy for refining species trees has the added advantage of complementing AnGST reconciliation process. High-frequency HGTs can be eliminated if an SPR between the transferred nodes (in the reverse direction of the HGT) is performed on the species tree. Consequently, GAnG generates a candidate species tree for each of the SPR moves associated with the 100 most frequently inferred HGT on the present species tree.

3.3.3 Rooting the species tree

Unlike typical supermatrix or supertree methods, GAnG does not require an outgroup sequence to infer a rooted species tree. Since the species tree root orients the direction of vertical inheritance at internal nodes, alternative rootings of the same unrooted species tree will have distinct AnGST scores when reconciled against a gene tree. Therefore, in addition to incorporating SPR moves, the candidate tree generation routine can also vary the position of the root node on the species tree.

3.3.4 Scoring a candidate species tree

A score S for a putative species tree topology T quantifies how well the species tree fits the set of gene trees according to the AnGST model:

$$S(T) = \sum_G \text{Rec}(G|T) \quad (3.1)$$

where $\text{Rec}(G|T)$ is the AnGST reconciliation score for a gene tree G , given the candidate species tree T . The AnGST model assigns a score of 3, 2, and 1 to each HGT, duplication, and loss event inferred, respectively [149].

3.4 Preliminary results

I have performed two preliminary tests using the GAnG algorithm, using simulated and real-world data generated from my previous study of microbial genome evolution [149]. These tests suggest:

1. The AnGST-based scoring function can discern between species trees of varying accuracy (Section 3.4.1)
2. HGTs can be used to propose SPR moves on the species tree that lead to lower AnGST reconciliation scores (Section 3.4.2).

3.4.1 AnGST scores increase with true species tree permutation

I evaluated how AnGST reconciliation scores changed as increasing phylogenetic noise was added to the Tree of Life [90], in order to test with real-world gene trees if the GAnG tree search process could be attracted towards the correct species tree. This experiment used 125 microbial gene families randomly selected from my previous study of microbial genome evolution [149]. A total of 100 sequenced genomes spanning all three domains of life were represented in these gene trees. Because the true species tree for the 100 sampled genomes is not known, I could not directly evaluate whether or not the GAnG algorithm could use the gene trees to recover the true species tree. Instead, I tested if the GAnG algorithm behaved in a manner consistent with less accurate species trees receiving higher reconciliation scores than more accurate species trees. I simulated a continuum of species tree accuracy by taking a topology (the Tree of Life) likely similar to the true tree, and randomly permuting it with between 1 and 10 SPR random moves. This procedure was used to generate 300 alternate species trees. Gene tree reconciliation scores against the species trees increased linearly with the number of SPRs performed on the Tree of Life (Figure 3.1; $R^2 = 0.67$), suggesting that more accurate species trees will receive lower reconciliation scores than less accurate species trees.

3.4.2 HGTs can be used to find a better-scoring tree

The GAnG algorithm was run to completion on a set of 12 sequenced archaeal species and 250 gene trees randomly chosen from my previous study on microbial genome evolution [149]. The initial species tree topology was pruned from the Ciccarelli & Bork Tree of Life (Fig. 3.2A) [90]. The algorithm terminated after four iterations, yielding a refined species tree whose reconciliations with gene trees were an average of 3.7% lower than the initial tree (Figure 3.2B). A single SPR move of the Crenarchaeota to a basal euryarchaeal lineage provided the most dramatic change in the refined topology, shifting *N. equitans* to a basal position on the archaeal tree.

3.5 Discussion

GAnG employs the AnGST reconciliation model to become the first GTP-based supertree method capable of accounting for HGT events and therefore appropriate for inferring prokaryotic species trees. Usage of the AnGST model carries two additional benefits. First, the AnGST algorithm includes a bootstrap amalgamation step that constructs a chimeric gene tree from the bootstrap subtrees that best conform to the species tree. As a result, gene families that evolved by vertical descent, but whose phylogenetic reconstructions are sensitive to sequence sampling variation, will be less likely to mislead the GTP process. Second, HGT events inferred by AnGST are directed between specific branches on the species tree and can therefore provide a heuristic guide for refining species trees. Frequent HGTs between two lineages may be the result of a topological error on the species tree and can be resolved by making these lineages sibling to one another on an updated tree.

Preliminary analyses of GAnG on simulated data demonstrate that the algorithm’s scoring function will be attracted towards correct species topologies (Fig. 3.1). The observed linear relationship between species tree score and tree randomization demonstrates also suggests that GAnG can accurately quantify relative differences in inaccuracy among trees. Thus, future implementations of GAnG’s tree search algorithm may be able to employ more sophisticated search algorithms, such as gradient descent.

The GAnG analysis on real-world data showed the algorithm capable of inferring a prokaryotic species tree largely in line with prior archaeal phylogenies, despite not relying on a “core” set of gene families free of HGT (Fig. 3.2). A single SPR move of the Crenarchaeota to a basal euryarchaeal lineage provided the most dramatic refinement to the starting Ciccarelli & Bork archaeal tree. This SPR shifted *N. equitans* from its initial sibling position with the Crenarchaeota (a position not supported by the archaeal phylogenetic literature) to a more basal position outgrouping both the Euryarchaeota and the Crenarchaeota. This outgroup location of *N. equitans* has been reported previously [155], but remains controversial [156]. The refined archaeal tree also includes a paraphyletic euryarchaeal clade that features the Thermoplas-

matales and Crenarchaeota as sibling to one another. A close relationship between the Thermoplasmatales and the Crenarchaeota has been reported by previous phylogenetic studies [157–159] and has been attributed to HGT between thermoplasma species and the Crenarchaeota [138].

Finally, these experiments both indicate that GAnG could be run on datasets composed of thousands of gene trees. The analysis of the archaeal tree with 250 gene trees took on the order of 3 days on a computer cluster. GAnG running time increases linearly with the number of gene trees, so analysis of a 1000 gene tree set would require approximately two weeks of compute time – a time-consuming experiment, but not prohibitively long. Moreover, a potential running time speedup is currently in development (see Future Directions below).

3.6 Future Directions

Future studies will investigate several potential improvements to the GAnG model, specifically in species tree scoring, running time improvement, and gene tree weighting. The benefits of each of these model changes will be investigated using gene and species tree simulation software that I have previously written during the development of AnGST [149].

A more accurate and more efficient version of GAnG will also be tested on a much larger archaeal dataset of 9053 gene families drawn from 70 sequenced archaeal genomes [160]. The resulting archaeal tree will be used to test hypotheses involving the basal position of the Korarchaeota and Thaumarchaeota on archaeal phylogenies and whether *N. equitans* represents a separate archaeal phylum.

3.6.1 Model improvements

Scoring modifications

The scoring function Eq. 3.1 may be biased by larger gene trees, which are likely to have a higher variance in reconciliation scores. An alternative scoring function robust to this bias is:

$$S(T) = \sum_G \text{Sign}(\text{Rec}(G|T) - \text{Rec}(G|R)) \quad (3.2)$$

This score is negative if there are more gene trees whose reconciliation scores are lower with the candidate tree than with the present species tree R . In this case, a candidate species tree would be accepted and become the new present species tree. Evidence that $S(T)$ will be positive for incorrect candidate trees is presented in Section 3.4.1 of the Preliminary Results. Summary of a tree’s fitness using the sign function also facilitates an algorithmic speedup described below in **Reducing running time**.

Alternatively, if we assume that reconciliation score variance grows with reconciliation score, exceptional changes in reconciliation score can be captured by the scoring

metric:

$$S(T) = \sum_G \text{Log} \left(\frac{\text{Rec}(G|T) - \text{Rec}(G|R)}{\text{Rec}(G|R)} \right) \quad (3.3)$$

In contrast to Equation 3.1, however, this scoring function may be biased by smaller gene trees, which are likely to have smaller reconciliation scores.

Reducing running time

I can avoid reconciling the entire set of gene trees if it can be quickly determined that a candidate species tree is less fit than the current species tree. I can perform this speedup using the scoring function in Equation 3.2 and by assuming that gene trees evolve independently from one another. Under this model, determining the sign of $S(T)$ can be likened to the problem of determining if a coin is fair using a finite number of coin tosses. To compute if there is a bias towards $[\text{Rec}(G|T) - \text{Rec}(G|R)]$ being positive or negative, we can count the total number of G for which this value is negative, N_- , or non-zero, N , and estimate the probability that a candidate species tree that changes the reconciliation score of G causes a lower reconciliation score:

$$p = \frac{N_-}{N} \quad (3.4)$$

However, p is an estimated probability with a standard error s_p

$$s_p = \sqrt{\frac{p(1-p)}{N}} \quad (3.5)$$

and a maximum error of

$$E = Z \times s_p \quad (3.6)$$

where Z is taken from a table of Z -values and associated confidence levels, calculated using a normal distribution [75].

If $p - E > 0.5$, the candidate topology can be accepted at the chosen confidence level. Similarly, if $p + E > 0.5$, the candidate topology can be rejected. Otherwise, the maximum error is too high to determine if the tree should be accepted or rejected

and more gene trees need to be reconciled. One way to determine the additional number of reconciliations to perform in this case would be to calculate $E = |0.5 - p|$, or the maximum error associated with p so that it can be determined if p is positive or negative. It can be shown that the number of reconciliations necessary to achieve this error, N_E , is equal to:

$$N_E = \frac{Z^2 p(1-p)}{E^2} \quad (3.7)$$

Gene tree weighting

Core gene approaches to species tree construction usually exclude gene families that show evidence of HGT [90]. However, the strict exclusion of gene families that carry only weak signals of HGT may be overly conservative. A single HGT event causes only one bifurcation on a gene tree to not be explained by vertical inheritance (i.e. a speciation event). A large gene tree with relatively few HGT can thus still be informative for phylogenomic purposes. This intuition can be incorporated into a scoring function by multiplying each gene tree score by a weighting factor $w(G)$. One method for calculating this weight would be to take the ratio of the number of HGT possible given a gene family's taxonomic distribution (which can be approximated by the square of the number of species L represented in the gene family), to the number of HGT inferred on the gene tree, $\text{HGT}(G)$:

$$w(G) = \frac{L(G)^2}{\text{HGT}(G)} \quad (3.8)$$

Translation-related gene families, which have been previously been relied upon for constructing archaeal phylogenies [138], are weighted more highly according to this scheme than the average gene family ($p < 10^{-30}$, Rank Sum test; Fig. 3.3). A caveat to this weighting scheme, however, is that $\text{HGT}(G)$ will change as the species tree is refined by GAnG.

I will also evaluate via simulation an alternative gene tree weighting function that uses gene tree branch lengths to downweight the influence of gene families suspected of

bearing transferred or duplicated genes. This approach relies on the assumption that genetic distances between gene copies descended from an HGT event will be shorter than expected, whereas genetic distances between certain gene copies descended from a duplication/loss scenarios will be longer than expected (see Figure 3.4 for an illustration of these phenomena). The following weighting function accounts for this evidence of non-vertical descent:

$$w(G) = 1 / \min_r \sum_{s_i, s_j} (rD(s_i, s_j|G) - D(s_i, s_j|R)) \quad (3.9)$$

This metric iterates over all pairs of species s_i and s_j represented in a gene tree G , and computes their pairwise distance on G using the function D . The difference between this distance and the expected distance (taken from the reference tree R) is then summed. To account for gene-specific rates of evolution, a rate term r is fit to G using a minimization function.

3.6.2 A tree of all sequenced archaea

Once I have completed optimizing the GAnG algorithm, I will construct an archaeal phylogeny using the arCOG dataset of 9504 archaeal gene families, which span 88% of sequenced archaeal genomes [160]. Due in part to the challenges of cultivating archaeal species [161], only 70 genomes have so far been sequenced from this domain of life; 4 of these genomes are the sole representatives of 3 of the 5 known archaeal phyla [155, 162–164] (Fig. 3.5). This uneven taxon sampling, along with suspected HGT events and unequal rates of lineage evolution, makes resolution of deep nodes on the archaeal phylogeny sensitive to the choice of analyzed genes. For example, a phylogeny built from a core set of 27 large and 23 small ribosomal subunit proteins supports a basal position of *N. equitans* on the archaeal tree, but the phylogeny of just the small subunit proteins suggests this species may only be a rapidly-evolving euryarchaeote [156].

An archaeal species tree built using the full arCOG dataset should correctly position *N. equitans*, unless phylogenetic artifacts systematically bias a large proportion

of genes in archaeal genomes. This tree may help resolve ongoing debates regarding the origins of other archaeal phyla. Previous studies with varying sets of core genes have placed the Korarchaeota either basal on the archaeal tree [165, 166], deep within the Crenarchaeota [162], or sibling to the Euryarchaeota [167]. Core gene studies have also reported conflicting positions for the Thaumarchaeota, positioning the phylum either within the Crenarchaeota [162] or basal on the archaeal tree [167, 168]. The GAnG algorithm offers a quantitative method for testing hypotheses of korarchaeal and thaumarchaeal origins against thousands of gene trees. Lessons learned from resolving these questions in archaeal history may also prove insightful for future efforts that use GAnG to reconstruct a larger three domain Tree of Life.

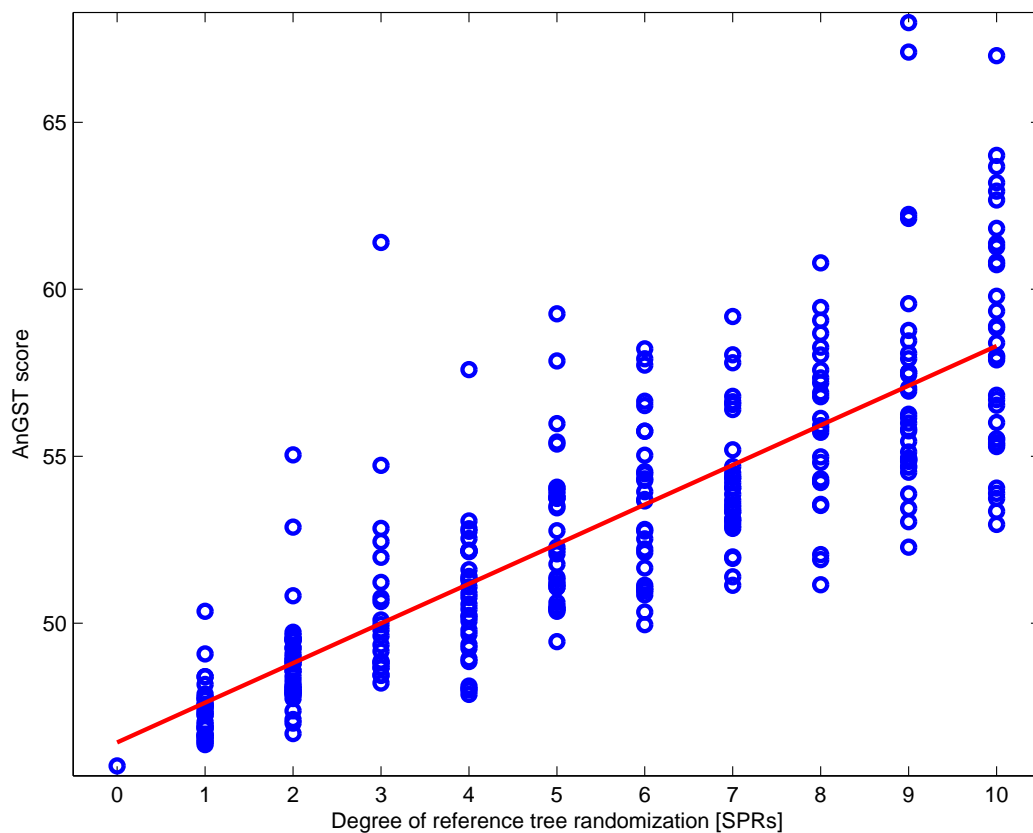


Figure 3.1: Average AnGST gene tree reconciliation scores as species tree randomization increases: Between 1-10 random SPR moves were made to 300 copies of the Tree of Life [90], to produce a continuum of species tree accuracy. Each of these species trees was reconciled against a set of 125 gene trees and the average AnGST score for each reconciliation is plotted on the y-axis. The R^2 of the best fit line (shown in red) to these data is 0.67.

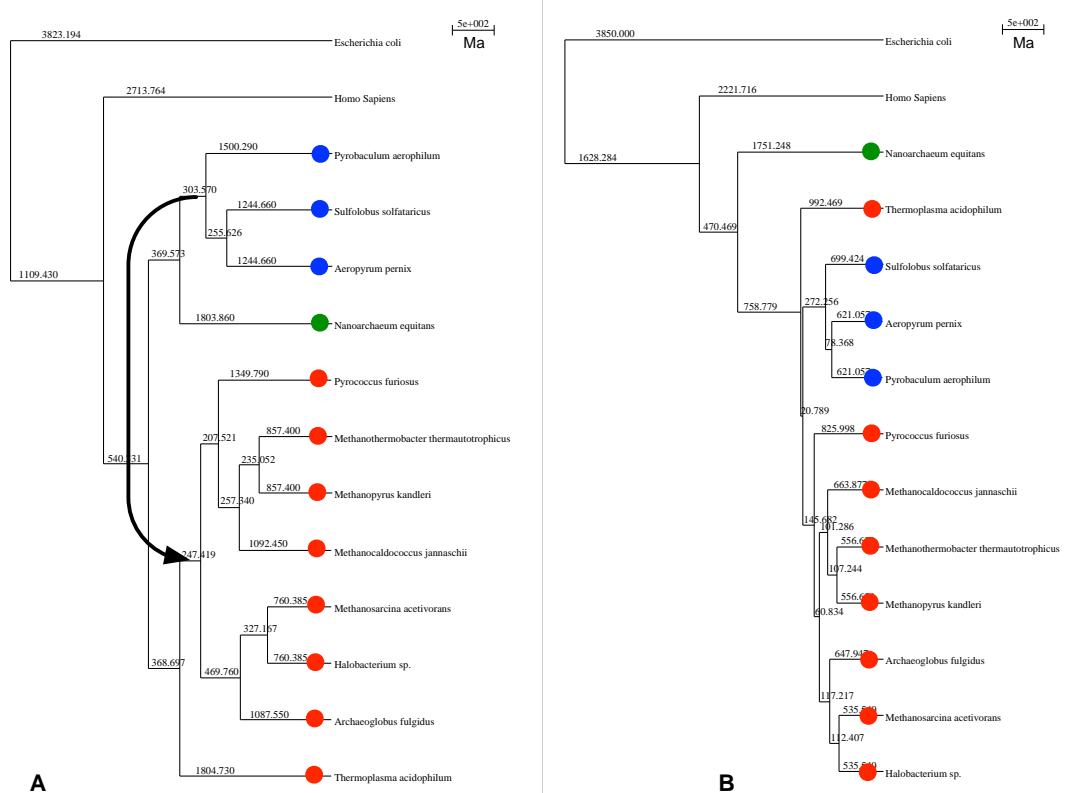


Figure 3.2: An archaeal phylogeny refined using 250 gene trees and HGT-proposed SPRs: (A) Initial phylogeny of archaeal species pruned from the Ciccarelli & Bork Tree of Life [90]. Crenarchaeal lineages are labeled in blue, euryarchaeal lineages are labeled in red, and the nanoarchaeal lineage is labeled in green. The most dramatic accepted SPR move on this topology is shown using a black arrow. (B) The resulting archaeal tree after 4 iterations of tree refinement. The average AnGST gene tree reconciliation score using this refined tree decreased 3.7% relative to the initial phylogeny.

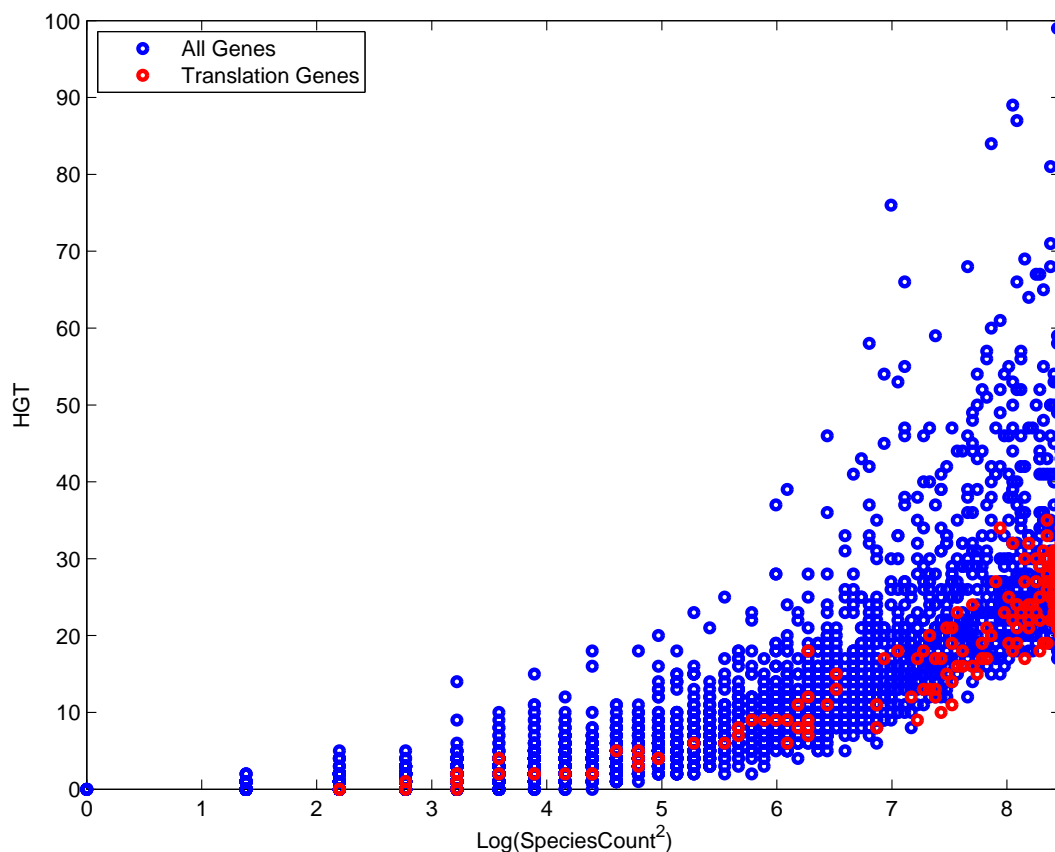


Figure 3.3: Inferred HGT as a function of the number of species represented in a gene tree: Each of the 9053 arCOG gene trees was reconciled against an archaeal species tree (Fig. 3.5). The number of inferred HGT for each gene family is plotted against the square of the number of species represented in the gene family (a rough approximation of the number of possible HGT) on a log-scale. The 201 gene families annotated by the arCOG database as involved in translation are shown in red, and all other gene families are shown in blue. Gene tree weights, as calculated by Eq. 3.8 are significantly higher for translation-associated gene families ($p < 10^{-30}$, Rank Sum test).

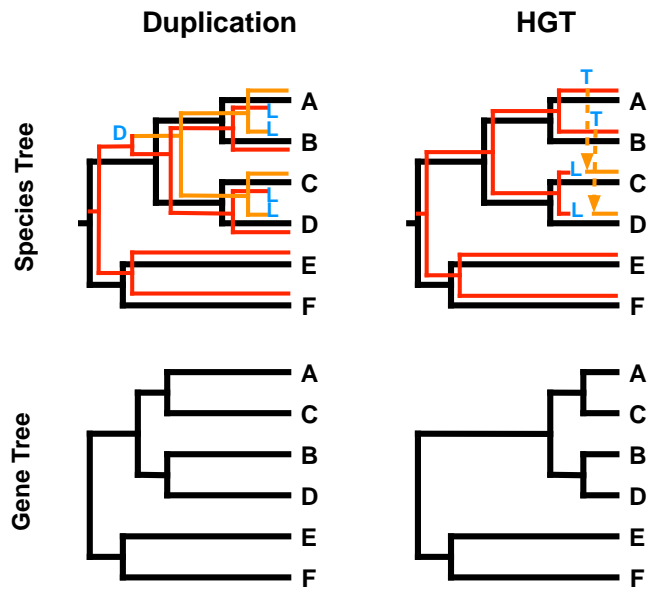


Figure 3.4: Expected gene tree branch lengths following duplication or HGT: Two hypothetical evolutionary histories are shown for a gene family. On the left, an ancestral duplication event (D), followed by four loss events (L), creates higher than expected genetic distance between species A and B, and between species C and D. On the right, HGT events (T) cause lower than expected genetic distance between species A and C, and between species B and D.

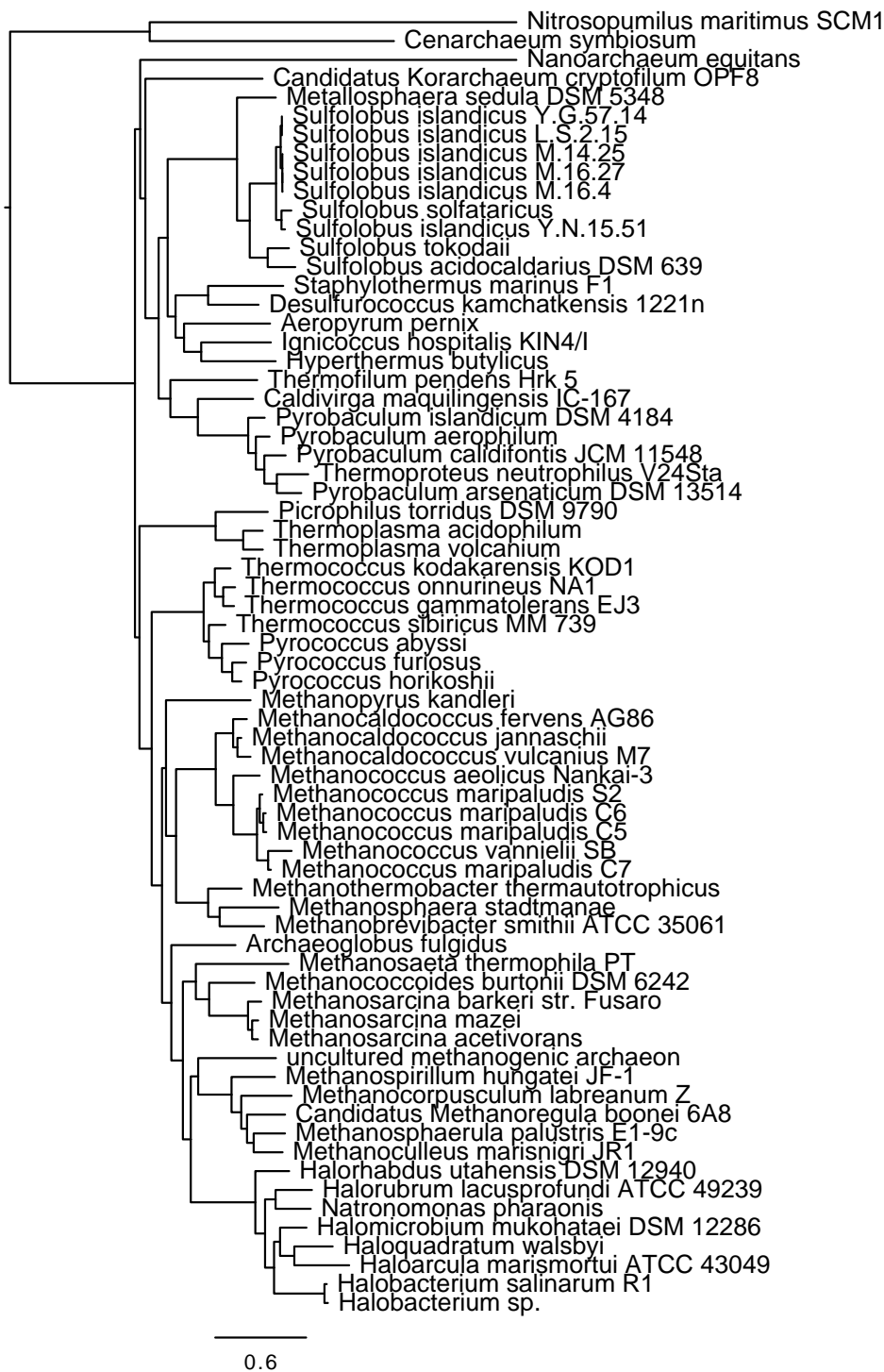


Figure 3.5: Maximum likelihood tree of archaeal species: This tree of 70 archaeal species will be the starting topology for the GAnG analysis of 9504 arCOG gene trees. The tree was built using a maximum likelihood analysis of 53 concatenated ribosomal proteins [138, 160].

Part III

Conclusion

This thesis contributes new algorithms for comparative microbial genomics, particularly in the subfields of microbial ecology and microbial genome evolution. In Chapter 1 of Part II, I presented the algorithm AdaptML, which can identify genetically- and ecologically-distinct bacterial populations. I used this tool to identify clusters of marine vibrio that appear to have differentiated in response to their nutrient preferences. Code for AdaptML has been made public and with the help of Albert Wang, an undergraduate researcher in Eric Alm’s laboratory, I have created a webserver where users can submit their own phylogenetic and ecological data for AdaptML to process online (<http://almlab.mit.edu/adaptml/>). Outside investigators have used these tools to run AnGST on their own datasets. Fred Cohan’s group ran the algorithm to find ecotypes in the genus *Bacillus* that would have been overlooked with traditional sequence divergence-based thresholds for calling species [11]. Oakley and colleagues used AdaptML to help confirm evidence for niche partitioning among sampled *Desulfobulbus* [169]. Other investigators have promoted using AdaptML for future studies of microbial evolution and ecology. Daniel Falush has suggested using AdaptML to infer the evolutionary history of microbe-host adaptation [170], and Ford Doolittle & Olga Zhaxybayeva have pointed out the algorithm’s advantages over strict threshold-based approaches to microbial ecology [171].

In Chapter 2, I introduced the algorithm AnGST, which reconciles an ultrametric reference tree and a gene tree to infer HGT, gene duplication, and gene family birth events in a chronological context. Implicit in these reconciliations are known constraints on organismal history drawn from paleontological and geochemical records. Analysis of 100 sequenced genomes with AnGST produced evidence for a massive expansion of microbial genetic diversity during the Archean eon, as well as the gradual oxygenation of the biosphere over the past 3 Ga. This later finding is in agreement with other studies that suggest secular changes in earth geochemistry were recorded in, and can now be mined from, microbial genomes [81, 83, 172, 173]. Further evidence of the link between the AnGST analysis and biogeochemistry could be uncovered by follow up studies on enzyme families whose first appearance can be dated using the sedimentary record. For example, Form 1 RuBisCO makes specific contacts with

molecular oxygen and first appeared 2.7-2.9 Ga according to carbon isotope fractionation results [174]. AnGST's predicted birth date for this enzyme's small subunit, between 2.0-3.0 Ga, is wide, but not incompatible with the fractionation results. Identification and analysis of other biomarkers whose age can be independently verified by AnGST will lead to potentially fruitful collaborations between genomicists, paleontologists, and biogeochemists.

Lastly, in Chapter 3, I introduced the supertree algorithm GAnG, which is capable of constructing prokaryotic species trees from thousands of gene trees. Preliminary GAnG analysis of 250 archaeal gene trees built from a subset of the sequenced archaeal genomes supports the hypothesis that Nanoarchaeota diverged from the last ancestor of the Archaea prior to the Crenarchaeota/Euryarchaeota split, but also appears sensitive to HGT from the Crenarchaeota to the Thermoplasma. Further improvements to the GAnG scoring function and gene tree weighting scheme are ongoing. An improved version of GAnG will be used to build a species tree relating 70 sequenced Archaea from all five known phyla. One important observation during that tree's construction will be the topology of the species tree scoring landscape near variations on deep node branching order. A relatively flat landscape in that regime may suggest that the phyla radiated so rapidly during early archaeal evolution that their branching order cannot be resolved [141] or that HGT dominated vertical descent during the formation of the archaeal phyla [175]. By contrast, a bowl-shaped landscape with a clear scoring minimum will provide support that a Tree of Life exists for the Archaea, and will encourage attempts to construct a universal Tree of Life with sequenced genomes from all three domains of life.

Part IV

Bibliography

Bibliography

- [1] Daniel Godoy, Gaynor Randle, Andrew J Simpson, David M Aanensen, Tyrone L Pitt, Reimi Kinoshita, and Brian G Spratt. Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J Clin Microbiol*, 41(5):2068–79, May 2003.
- [2] Fergus G Priest, Margaret Barker, Les W J Baillie, Edward C Holmes, and Martin C J Maiden. Population structure and evolution of the *Bacillus cereus* group. *J Bacteriol*, 186(23):7959–70, Dec 2004.
- [3] William P Hanage, Christophe Fraser, and Brian G Spratt. Fuzzy species among recombinogenic bacteria. *BMC Biol*, 3:6, Jan 2005.
- [4] Dirk Gevers, Frederick M Cohan, Jeffrey G Lawrence, Brian G Spratt, Tom Coenye, Edward J Feil, Erko Stackebrandt, Yves Van de Peer, et al. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol*, 3(9):733–9, Sep 2005.
- [5] Frederick M Cohan and Elizabeth B Perry. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol*, 17(10):R373–86, May 2007.
- [6] Christophe Fraser, William P Hanage, and Brian G Spratt. Recombination and the nature of bacterial speciation. *Science*, 315(5811):476–80, Jan 2007.
- [7] B. Jesse Shapiro, Jonathan Friedman, Otto X. Cordero, Sarah Preheim, Sonia Timberlake, Martin Polz, and Eric Alm. Recombination in the core and flexible genome drives ecological differentiation in sympatric ocean microbes. *In progress*.
- [8] Andrew P Martin. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol*, 68(8):3673–82, Aug 2002.
- [9] Catherine Lozupone and Rob Knight. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, 71(12):8228–35, Dec 2005.
- [10] Alexander Koeppel, Elizabeth B Perry, Johannes Sikorski, Danny Krizanc, Andrew Warner, David M Ward, Alejandro P Rooney, Evelyn Brambila, et al. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci USA*, 105(7):2504–9, Feb 2008.
- [11] Nora Connor, Johannes Sikorski, Alejandro P Rooney, Sarah Kopac, Alexander F Koeppel, Andrew Burger, Scott G Cole, Elizabeth B Perry, et al. Ecology of speciation in the genus *Bacillus*. *Appl Environ Microbiol*, 76(5):1349–58, Mar 2010.
- [12] H Ochman, J G Lawrence, and E A Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, May 2000.
- [13] Eugene V Koonin. Darwinian evolution in the light of genomics. *Nucleic Acids Res*, 37(4):1011–34, Mar 2009.
- [14] N A Moran and A Mira. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*, 2(12), Jan 2001.
- [15] M M Riehle, A F Bennett, and A D Long. Genetic architecture of thermal adaptation

- in *Escherichia coli*. *Proc Natl Acad Sci USA*, 98(2):525–30, Jan 2001.
- [16] Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–24, Apr 2004.
- [17] Nancy A Moran and Tyler Jarvik. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science*, 328(5978):624–7, Apr 2010.
- [18] Mary E Rumpho, Jared M Worful, Jungho Lee, Krishna Kannan, Mary S Tyler, Debashish Bhattacharya, Ahmed Moustafa, and James R Manhart. Horizontal gene transfer of the algal nuclear gene *psbO* to the photosynthetic sea slug *Elysia chlorotica*. *Proc Natl Acad Sci USA*, 105(46):17867–71, Nov 2008.
- [19] Julie C Dunning Hotopp, Michael E Clark, Deodoro C S G Oliveira, Jeremy M Foster, Peter Fischer, Mónica C Muñoz Torres, Jonathan D Giebel, Nikhil Kumar, et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, 317(5845):1753–6, Sep 2007.
- [20] Natsuko Kondo, Naruo Nikoh, Nobuyuki Ijichi, Masakazu Shimada, and Takema Fukatsu. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci USA*, 99(22):14280–5, Oct 2002.
- [21] J G Lawrence and H Ochman. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA*, 95(16):9413–7, Aug 1998.
- [22] Yoji Nakamura, Takeshi Itoh, Hideo Matsuda, and Takashi Gojobori. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet*, 36(7):760–6, Jul 2004.
- [23] Dirk Gevers, Klaas Vandepoele, Cedric Simillon, and Yves Van de Peer. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol*, 12(4):148–54, Apr 2004.
- [24] Eric Alm, Katherine Huang, and Adam Arkin. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol*, 2(11):e143, Nov 2006.
- [25] Victor Kunin and Christos A Ouzounis. The balance of driving forces during genome evolution in prokaryotes. *Genome Res*, 13(7):1589–94, Jul 2003.
- [26] Boris G Mirkin, Trevor I Fenner, Michael Y Galperin, and Eugene V Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, 3:2, Jan 2003.
- [27] Berend Snel, Peer Bork, and Martijn A Huynen. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, 12(1):17–25, Jan 2002.
- [28] Olga Zhaxybayeva, J Peter Gogarten, Robert L Charlebois, W Ford Doolittle, and R Thane Papke. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res*, 16(9):1099–108, Sep 2006.
- [29] Vincent Daubin, Nancy A Moran, and Howard Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301(5634):829–32, Aug 2003.
- [30] R D Page and M A Charleston. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, 7(2):231–40, Apr 1997.
- [31] M A Charleston. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149(2):191–223, May 1998.
- [32] Matthew W Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol*, 8(7):R141, Jan 2007.
- [33] Matthew D Rasmussen and Manolis Kellis. Accurate gene-tree reconstruction by

- learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*, 17(12):1932–42, Dec 2007.
- [34] Orjan Akerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA*, 106(14):5714–9, Apr 2009.
- [35] Patrick J Keeling and Jeffrey D Palmer. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, 9(8):605–18, Aug 2008.
- [36] J Peter Gogarten and Jeffrey P Townsend. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, 3(9):679–87, Sep 2005.
- [37] Stephen J Giovannoni and Ulrich Stingl. Molecular diversity and ecology of microbial plankton. *Nature*, 437(7057):343–8, Sep 2005.
- [38] Edward F DeLong, Christina M Preston, Tracy Mincer, Virginia Rich, Steven J Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matthew B Sullivan, et al. Community genomics among stratified microbial assemblages in the ocean’s interior. *Science*, 311(5760):496–503, Jan 2006.
- [39] Martin F Polz, Dana E Hunt, Sarah P Preheim, and Daniel M Weinreich. Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philos Trans R Soc Lond, B, Biol Sci*, 361(1475):2009–21, Nov 2006.
- [40] Alban Ramette and James M Tiedje. Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proc Natl Acad Sci USA*, 104(8):2761–6, Feb 2007.
- [41] Jennifer B Hughes Martiny, Brendan J M Bohannan, James H Brown, Robert K Colwell, Jed A Fuhrman, Jessica L Green, M Claire Horner-Devine, Matthew Kane, et al. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol*, 4(2):102–12, Feb 2006.
- [42] Silvia G Acinas, Vanja Klepac-Ceraj, Dana E Hunt, Chanathip Pharino, Ivica Ceraj, Daniel L Distel, and Martin F Polz. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430(6999):551–4, Jul 2004.
- [43] Zackary I Johnson, Erik R Zinser, Allison Coe, Nathan P McNulty, E Malcolm S Woodward, and Sallie W Chisholm. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science*, 311(5768):1737–40, Mar 2006.
- [44] Mike J Ferris, Michael Köhl, Andrea Wieland, and David M Ward. Cyanobacterial ecotypes in different optical microenvironments of a 68 degrees C hot spring mat community revealed by 16S-23S rRNA internal transcribed spacer region variation. *Appl Environ Microbiol*, 69(5):2893–8, May 2003.
- [45] W Ford Doolittle and R Thane Papke. Genomics and the bacterial species problem. *Genome Biology*, 7(9):116, Jan 2006.
- [46] Adam C Retchless and Jeffrey G Lawrence. Temporal fragmentation of speciation in bacteria. *Science*, 317(5841):1093–6, Aug 2007.
- [47] J R Thompson and M F Polz. *The Biology of Vibrios*. ASM Press, Washington, DC, 2006.
- [48] Thomas Kjørboe, Hans-Peter Grossart, Helle Ploug, and Kam Tang. Mechanisms and rates of bacterial colonization of sinking aggregates. *Appl Environ Microbiol*, 68(8):3996–4006, Aug 2002.
- [49] L Pomeroy, R Hanson, P McGillivray, B Sherr, D Kirchman, and D Deibel. Microbiology and Chemistry of Fecal Products of Pelagic Tunicates: Rates and Fates. *Bull. Mar. Sci.*, 35(3):426–439, 1984.
- [50] C Panagiotopoulos, R Sempéré, and I Obernosterer. Bacterial degradation of large particles in the southern Indian Ocean using in vitro incubation experiments. *Organic Geochemistry*, 33:985–1000, Jan 2002.

- [51] J F Heidelberg, K B Heidelberg, and R R Colwell. Bacteria of the gamma-subclass Proteobacteria associated with zooplankton in Chesapeake Bay. *Appl Environ Microbiol*, 68(11):5498–507, Nov 2002.
- [52] Dana E Hunt, Dirk Gevers, Nisha M Vahora, and Martin F Polz. Conservation of the chitin utilization pathway in the Vibrionaceae. *Appl Environ Microbiol*, 74(1):44–51, Jan 2008.
- [53] Jakob Pernthaler and Rudolf Amann. Fate of heterotrophic microbes in pelagic habitats: focus on populations. *Microbiol Mol Biol Rev*, 69(3):440–61, Sep 2005.
- [54] Janelle R Thompson, Sarah Pacocha, Chanathip Pharino, Vanja Klepac-Ceraj, Dana E Hunt, Jennifer Benoit, Ramahi Sarma-Rupavtarm, Daniel L Distel, et al. Genotypic diversity within a natural coastal bacterioplankton population. *Science*, 307(5713):1311–3, Feb 2005.
- [55] W Maddison and M Slatkin. Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution*, 45(5):1184–1197, Jan 1991.
- [56] Fredrik Ronquist. Bayesian inference of character evolution. *Trends Ecol Evol (Amst)*, 19(9):475–81, Sep 2004.
- [57] Janelle R Thompson, Mark A Randa, Luisa A Marcelino, Aoy Tomita-Mitchell, Eelin Lim, and Martin F Polz. Diversity and dynamics of a north atlantic coastal Vibrio community. *Appl Environ Microbiol*, 70(7):4103–10, Jul 2004.
- [58] Vincent Daubin and Nancy A Moran. Comment on "The origins of genome complexity". *Science*, 306(5698):978, Nov 2004.
- [59] Frederick M Cohan. Sexual isolation and speciation in bacteria. *Genetica*, 116(2-3):359–70, Nov 2002.
- [60] J Majewski. Sexual isolation in bacteria. *FEMS Microbiol Lett*, 199(2):161–9, May 2001.
- [61] C von Mering, P Hugenholtz, J Raes, S G Tringe, T Doerks, L J Jensen, N Ward, and P Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–30, Feb 2007.
- [62] S G Acinas, J Antón, and F Rodríguez-Valera. Diversity of free-living and attached bacteria in offshore Western Mediterranean waters as depicted by analysis of genes encoding 16S rRNA. *Appl Environ Microbiol*, 65(2):514–22, Feb 1999.
- [63] B C Crump, E V Armbrust, and J A Baross. Phylogenetic analysis of particle-associated and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean. *Appl Environ Microbiol*, 65(7):3192–204, Jul 1999.
- [64] L Riemann and A Winding. Community Dynamics of Free-living and Particle-associated Bacterial Assemblages during a Freshwater Phytoplankton Bloom. *Microbial Ecology*, 42(3):274–285, Oct 2001.
- [65] N Selje and M Simon. Composition and dynamics of particle-associated and free-living bacterial communities in the Weser estuary, Germany. *Aquatic Microbial Ecology*, 30:221–237, Jan 2003.
- [66] E Delong, D Franks, and A Alldredge. Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography*, 38(5):924–934, Jan 1993.
- [67] S H Goh, S Potter, J O Wood, S M Hemmingsen, R P Reynolds, and A W Chow. HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. *J Clin Microbiol*, 34(4):818–23, Apr 1996.
- [68] Erko Stackebrandt and M Goodfellow, editors. *Nucleic acid techniques in bacterial systematics*. Wiley and Sons, Chichester, UK, 1991.
- [69] J R Cole, B Chai, R J Farris, Q Wang, A S Kulam-Syed-Mohideen, D M McGarrell, A M Bandela, E Cardenas, et al. The

- ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*, 35(Database issue):D169–72, Jan 2007.
- [70] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 1990.
- [71] Scott R Santos and Howard Ochman. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environmental Microbiology*, 6(7):754–9, Jul 2004.
- [72] Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, Oct 2003.
- [73] M Hasegawa, Y Iida, T Yano, F Takaiwa, and M Iwabuchi. Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. *J Mol Evol*, 22(1):32–8, Jan 1985.
- [74] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8, Jan 2007.
- [75] John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press: Belmont, CA, 1994.
- [76] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Inc., Sunderland, MA, 2003.
- [77] Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, 1974.
- [78] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme J. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
- [79] E G Nisbet and N H Sleep. The habitat and nature of early life. *Nature*, 409(6823):1083–91, Feb 2001.
- [80] Birger Rasmussen, Ian R Fletcher, Jochen J Brocks, and Matt R Kilburn. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, 455(7216):1101–4, Oct 2008.
- [81] Christopher L Dupont, Song Yang, Brian Palenik, and Philip E Bourne. Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc Natl Acad Sci USA*, 103(47):17822–7, Nov 2006.
- [82] M Saito, D Sigman, and F Morel. The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean-Proterozoic boundary? *Inorganica Chimica Acta*, 356:308–318, Jan 2003.
- [83] A Zerkle, C House, and S Brantley. Biogeochemical signatures through time as inferred from whole microbial genomes. *American Journal of Science*, 305:467–502, Jan 2005.
- [84] D Yang, Y Oyaizu, H Oyaizu, G J Olsen, and C R Woese. Mitochondrial origins. *Proc Natl Acad Sci USA*, 82(13):4443–7, Jul 1985.
- [85] S J Giovannoni, S Turner, G J Olsen, S Barns, D J Lane, and N R Pace. Evolutionary relationships among cyanobacteria and green chloroplasts. *J Bacteriol*, 170(8):3584–92, Aug 1988.
- [86] D J De Marais. Evolution. When did photosynthesis emerge on Earth? *Science*, 289(5485):1703–5, Sep 2000.
- [87] J Gogarten, W Doolittle, and Jeffrey Lawrence. Prokaryotic Evolution in Light of Gene Transfer. *Mol Biol Evol*, 19(12):2226, Dec 2002.
- [88] R Jain, M C Rivera, and J A Lake. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA*, 96(7):3801–6, Mar 1999.
- [89] Mark A Ragan and Robert G Beiko. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond, B, Biol Sci*, 364(1527):2241–51, Aug 2009.

- [90] Francesca D Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–7, Mar 2006.
- [91] Thomas Lepage, David Bryant, Hervé Philippe, and Nicolas Lartillot. A general comparison of relaxed molecular clock models. *Mol Biol Evol*, 24(12):2669–80, Dec 2007.
- [92] S J Mojzsis, G Arrhenius, K D McKee-gan, T M Harrison, A P Nutman, and C R Friend. Evidence for life on Earth before 3,800 million years ago. *Nature*, 384(6604):55–9, Nov 1996.
- [93] M Rosing. ^{13}C -Depleted carbon microparticles in 3700-Ma sea-floor sedimentary rocks from west greenland. *Science*, 283(5402):674–6, Jan 1999.
- [94] Jessica Garvin, Roger Buick, Ariel D Anbar, Gail L Arnold, and Alan J Kaufman. Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. *Science*, 323(5917):1045–8, Feb 2009.
- [95] Ariel D Anbar, Yun Duan, Timothy W Lyons, Gail L Arnold, Brian Kendall, Robert A Creaser, Alan J Kaufman, Gwyneth W Gordon, et al. A whiff of oxygen before the great oxidation event? *Science*, 317(5846):1903–6, Sep 2007.
- [96] Alan J Kaufman, David T Johnston, James Farquhar, Andrew L Masterson, Timothy W Lyons, Steve Bates, Ariel D Anbar, Gail L Arnold, et al. Late Archean biospheric oxygenation and atmospheric evolution. *Science*, 317(5846):1900–3, Sep 2007.
- [97] Christopher T Reinhard, Rob Raiswell, Clint Scott, Ariel D Anbar, and Timothy W Lyons. A late Archean sulfidic sea stimulated by early oxidative weathering of the continents. *Science*, 326(5953):713–6, Oct 2009.
- [98] J J Brocks, G A Logan, R Buick, and R E Summons. Archean molecular fossils and the early rise of eukaryotes. *Science*, 285(5430):1033–6, Aug 1999.
- [99] Jacob R Waldbauer, Laura S Sherman, Dawn Y Sumner, and Roger E Summons. Late Archean molecular fossils from the Transvaal Supergroup record the antiquity of microbial diversity and aerobiosis. *Pre-cambrian Research*, 169(1-4):28–47, Dec 2009.
- [100] S Golubic, V N Sergeev, and A H Knoll. Mesoproterozoic Archaeoellipsoides: akinetes of heterocystous cyanobacteria. *Lethaia*, 28:285–98, Jan 1995.
- [101] N Butterfield. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/ Neoproterozoic radiation of eukaryotes. *Paleobiology*, 26(3):386–404, Jan 2000.
- [102] Gordon D Love, Emmanuelle Grosjean, Charlotte Stalvies, David A Fike, John P Grotzinger, Alexander S Bradley, Amy E Kelly, Maya Bhatia, et al. Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature*, 457(7230):718–21, Feb 2009.
- [103] Edward B Daeschler, Neil H Shubin, and Farish A Jenkins. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature*, 440(7085):757–63, Apr 2006.
- [104] E Daeschler, N Shubin, K Thomson, and W Amaral. A Devonian Tetrapod from North America. *Science*, 265(5172):639–642, Jul 1994.
- [105] N Moran, M Munson, P Baumann, and H Ishikawa. A Molecular Clock in Endosymbiotic Bacteria is Calibrated Using the Insect Hosts. *Proceedings of the Royal Society B: Biological Sciences*, 253:167–171, Jan 1993.
- [106] Lars Juhl Jensen, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*, 36(Database issue):D250–4, Jan 2008.
- [107] J G Jelesko, R Harper, M Furuya, and W Gruissem. Rare germinal unequal

- crossing-over leading to recombinant gene formation and gene duplication in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*, 96(18):10302–7, Aug 1999.
- [108] Michael Lynch and John S Conery. The origins of genome complexity. *Science*, 302(5649):1401–4, Nov 2003.
- [109] M Sugiura, T Hirose, and M Sugita. Evolution and mechanism of translation in chloroplasts. *Annu Rev Genet*, 32:437–59, Jan 1998.
- [110] D Fischer and D Eisenberg. Finding families for genomic ORFans. *Bioinformatics*, 15(9):759–62, Sep 1999.
- [111] C Scott, T W Lyons, A Bekker, Y Shen, S W Poulton, X Chu, and A D Anbar. Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature*, 452(7186):456–9, Mar 2008.
- [112] Kurt O Konhauser, Ernesto Pecoits, Stefan V Lalonde, Dominic Papineau, Euan G Nisbet, Mark E Barley, Nicholas T Arndt, Kevin Zahnle, et al. Oceanic nickel depletion and a methanogen famine before the Great Oxidation Event. *Nature*, 458(7239):750–3, Apr 2009.
- [113] D Canfield. A new model for Proterozoic ocean chemistry. *Nature*, 396(6710):450–453, Jan 1998.
- [114] A Bekker, H D Holland, P-L Wang, D Rumble, H J Stein, J L Hannah, L L Coetzee, and N J Beukes. Dating the rise of atmospheric oxygen. *Nature*, 427(6970):117–20, Jan 2004.
- [115] D Canfield. The early history of atmospheric oxygen: homage to Robert M. Garrels. *Annu. Rev. Earth Planet. Sci.*, 33:1–36, Jan 2005.
- [116] Michael J Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–2, Jan 2003.
- [117] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.
- [118] Eric J Alm, Katherine H Huang, Morgan N Price, Richard P Koche, Keith Keller, Inna L Dubchak, and Adam P Arkin. The MicrobesOnline Web site for comparative genomics. *Genome Res*, 15(7):1015–22, Jul 2005.
- [119] JP Huelsenbeck, B Rannala, and B Larget. *Tangled Trees: Phylogenies, Cospeciation, and Coevolution*. University of Chicago Press, Chicago, 2002.
- [120] L Arvestad, A Berglund, J Lagergren, and B Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB’04*, pages 326–335, Jan 2004.
- [121] Dave MacLeod, Robert L Charlebois, Ford Doolittle, and Eric Baptiste. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol*, 5(1):27, Jan 2005.
- [122] J Bergsten. A review of long-branch attraction. *Cladistics*, 21:163–193, Jan 2005.
- [123] A Rambaut and N Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, Jan 1997.
- [124] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.
- [125] Tal Dagan and William Martin. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA*, 104(3):870–5, Jan 2007.
- [126] Nicolas Lartillot and Hervé Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21(6):1095–109, Jun 2004.
- [127] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, et al. The COG

- database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.
- [128] Gerard Talavera and Jose Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4):564–77, Aug 2007.
- [129] J Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, 17(4):540–52, Apr 2000.
- [130] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7, Jan 2004.
- [131] Frédéric Delsuc, Henner Brinkmann, and Hervé Philippe. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, 6(5):361–75, May 2005.
- [132] Fengrong Ren, Hiroshi Tanaka, and Ziheng Yang. A likelihood look at the supermatrix-supertree controversy. *Gene*, 441(1-2):119–25, Jul 2009.
- [133] J Slowinski and R Page. How should species phylogenies be inferred from sequence data? *Syst Biol*, 48(4):814–825, Jan 1999.
- [134] Eric Baptiste, Henner Brinkmann, Jennifer A Lee, Dorothy V Moore, Christoph W Sensen, Paul Gordon, Laure Duruflé, Terry Gaasterland, et al. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci USA*, 99(3):1414–9, Feb 2002.
- [135] J Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Biol*, 27(4):401–410, Jan 1978.
- [136] P J Lockhart, M A Steel, M D Hendy, and D Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*, 11(4):605–12, Jul 1994.
- [137] Matthew J Phillips, Frédéric Delsuc, and David Penny. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*, 21(7):1455–8, Jul 2004.
- [138] Oriane Matte-Tailliez, Céline Brochier, Patrick Forterre, and Hervé Philippe. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol*, 19(5):631–9, May 2002.
- [139] Tal Dagan and William Martin. The tree of one percent. *Genome Biol*, 7(10):118, Jan 2006.
- [140] Vincent Daubin and Howard Ochman. Quartet mapping and the extent of lateral transfer in bacterial genomes. *Mol Biol Evol*, 21(1):86–9, Jan 2004.
- [141] Pere Puigbò, Yuri I Wolf, and Eugene V Koonin. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *Journal of Biology*, 8(6):59, Jan 2009.
- [142] Hervé Philippe, Elizabeth A Snell, Eric Baptiste, Philippe Lopez, Peter W H Holland, and Didier Casane. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol*, 21(9):1740–52, Sep 2004.
- [143] Hervé Philippe and Christophe J Douady. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol*, 6(5):498–505, Oct 2003.
- [144] Olaf R P Bininda-Emonds. The evolution of supertrees. *Trends Ecol Evol (Amst)*, 19(6):315–22, Jun 2004.
- [145] M A Ragan. Phylogenetic inference based on matrix representation of trees. *Mol Phylogenet Evol*, 1(1):53–8, Mar 1992.
- [146] J B Slowinski, A Knight, and A P Rooney. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Mol Phylogenet Evol*, 8(3):349–62, Dec 1997.
- [147] R D Page. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14(9):819–20, Jan 1998.

- [148] A Wehe, M Bansal, J Burleigh, and O Eulenstein. DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, May 2008.
- [149] LA David and EJ Alm. Rapid evolutionary innovation during an Archean Genetic Expansion. *Submitted*, 2010.
- [150] M Bansal, J Burleigh, O Eulenstein, and A Wehe. Heuristics for the gene-duplication problem: A (n) speed-up for the local search. *RECOMB'07*, pages 238–252, Jan 2007.
- [151] L Billera, S Holmes, and K Vogtmann. Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27(4):733–767, Jan 2001.
- [152] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59(3):307–21, May 2010.
- [153] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, Jan 2010.
- [154] A Stamatakis, T Ludwig, and H Meier. RAXML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–63, Feb 2005.
- [155] Elizabeth Waters, Michael J Hohn, Ivan Ahel, David E Graham, Mark D Adams, Mary Barnstead, Karen Y Beeson, Lisa Bibbs, et al. The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. *Proc Natl Acad Sci USA*, 100(22):12984–8, Oct 2003.
- [156] Celine Brochier, Simonetta Gribaldo, Yvan Zivanovic, Fabrice Confalonieri, and Patrick Forterre. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biology*, 6(5):R42, Jan 2005.
- [157] Yuri I Wolf, Igor B Rogozin, Nick V Grishin, and Eugene V Koonin. Genome trees and the tree of life. *Trends Genet*, 18(9):472–9, Sep 2002.
- [158] Alexei I Slesarev, Katja V Mezhevaya, Kira S Makarova, Nikolai N Polushin, Olga V Shcherbinina, Vera V Shakhova, Galina I Belova, L Aravind, et al. The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proc Natl Acad Sci USA*, 99(7):4644–9, Apr 2002.
- [159] Ji Qi, Bin Wang, and Bai-lin Hao. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol*, 58(1):1–11, Jan 2004.
- [160] Kira S Makarova, Alexander V Sorokin, Pavel S Novichkov, Yuri I Wolf, and Eugene V Koonin. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct*, 2:33, Jan 2007.
- [161] Ken Neelson. A Korarchaeote yields to genome sequencing. *Proc Natl Acad Sci USA*, 105(26):8805–6, Jul 2008.
- [162] James G Elkins, Mircea Podar, David E Graham, Kira S Makarova, Yuri Wolf, Lennart Randau, Brian P Hedlund, Céline Brochier-Armanet, et al. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci USA*, 105(23):8102–7, Jun 2008.
- [163] Steven J Hallam, Konstantinos T Konstantinidis, Nik Putnam, Christa Schleper, Yoh ichi Watanabe, Junichi Sugahara, Christina Preston, José de la Torre, et al. Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. *Proc Natl Acad Sci USA*, 103(48):18296–301, Nov 2006.
- [164] Harald Huber, Michael J Hohn, Reinhard Rachel, Tanja Fuchs, Verena C Wimmer, and Karl O Stetter. A new phylum of Archaea represented by a nano-sized hyperthermophilic symbiont. *Nature*, 417(6884):63–7, May 2002.

- [165] S M Barns, C F Delwiche, J D Palmer, and N R Pace. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc Natl Acad Sci USA*, 93(17):9188–93, Aug 1996.
- [166] Thomas A Auchtung, Cristina D Takacs-Vesbach, and Colleen M Cavanaugh. 16S rRNA phylogenetic investigation of the candidate division "Korarchaeota". *Appl Environ Microbiol*, 72(7):5077–82, Jul 2006.
- [167] Anja Spang, Roland Hatzenpichler, Céline Brochier-Armanet, Thomas Rattei, Patrick Tischler, Eva Spieck, Wolfgang Streit, David A Stahl, et al. Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends Microbiol*, 18(8):331–340, Aug 2010.
- [168] Céline Brochier-Armanet, Bastien Bous-sau, Simonetta Gribaldo, and Patrick Forterre. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol*, 6(3):245–52, Mar 2008.
- [169] Brian B Oakley, Franck Carbonero, Christopher J van der Gast, Robert J Hawkins, and Kevin J Purdy. Evolutionary divergence and biogeography of sympatric niche-differentiated bacterial populations. *ISME J*, 4(4):488–97, Apr 2010.
- [170] Daniel Falush. Toward the use of genomics to study microevolutionary change in bacteria. *PLoS Genet*, 5(10):e1000627, Oct 2009.
- [171] W Doolittle and O Zhaxybayeva. Metagenomics and the Units of Biological Organization. *BioScience*, 60(2):102–112, Jan 2010.
- [172] Christopher L Dupont, Andrew Butcher, Ruben E Valas, Philip E Bourne, and Gustavo Caetano-Anollés. History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc Natl Acad Sci USA*, 107(23):10567–72, Jun 2010.
- [173] Jason Raymond and Daniel Segrè. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311(5768):1764–7, Mar 2006.
- [174] E G Nisbet, N V Grassineau, C J Howe, P I Abell, M Regelous, and R E R Nisbet. The age of Rubisco: the evolution of oxygenic photosynthesis. *Geobiology*, 5:311–335, Jan 2007.
- [175] Eugene V Koonin. The Biological Big Bang model for the major transitions in evolution. *Biol Direct*, 2:21, Jan 2007.